EXTRA

MUNDO CIENTIFICO

Extra • 995 ptas.

La Recherche

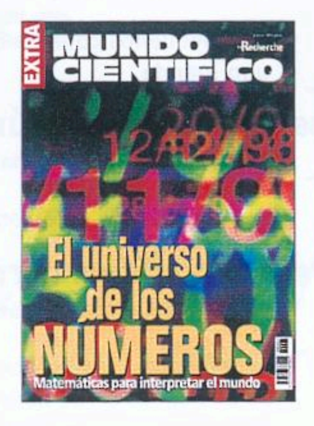
El universo de los NÚMEROS

Matemáticas para interpretar el mundo

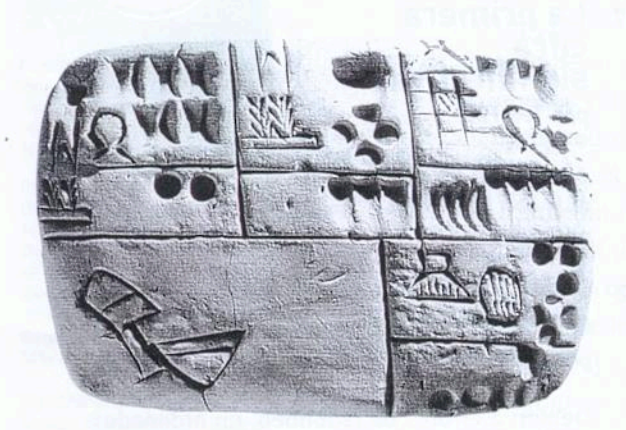


sumario

PORTADA
«¿Qué es un número? Mientras formulaba la pregunta me di cuenta de que no conocía la respuesta». Philip I. Davis, *El universo matemático*, 1982.
La vieja pregunta es: ¿cabe definir un número?







12. El nacimiento del número
Tablilla de contabilidad de raciones (Uruk, hacia 3000 a.C.). Marcas impresas en arcilla.



24. ¿Existen los números infinitos?
¿Puede ser un infinito mayor que otro? ¿Tiene realidad el infinito, o es una simple «ficción útil» del cálculo, como pensaba Leibniz?

Todo lo que se puede conocer tiene un número. ¿Podría ser que el número fuese una realidad en cierta manera superior, que no sólo preexistiría a la escritura, sino también al hombre e incluso a los elementos.

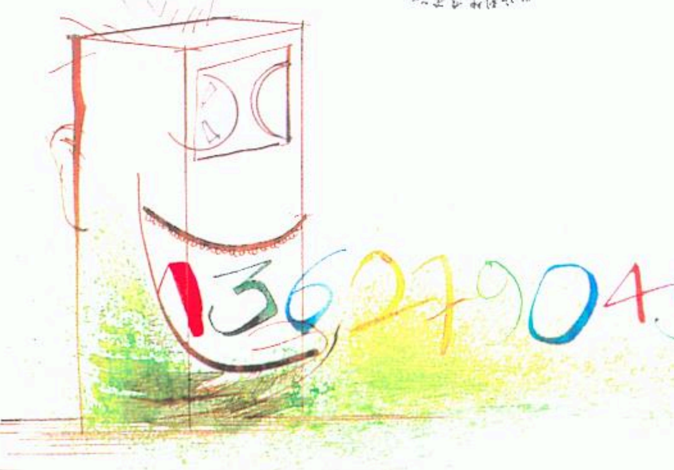


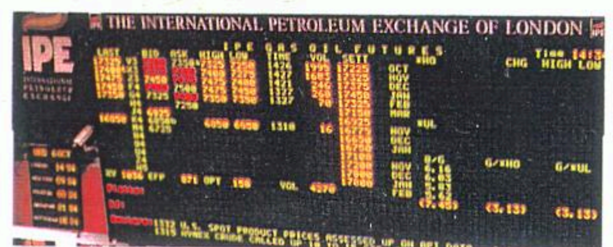


61. La doble corrección
Los turbocódigos se utilizarán en las futuras misiones remotas de las agencias NASA (norteamericana) y ESA (europea).

72. Los códigos correctores
Uno de los códigos binarios más antiguos es sin duda el constituido por los 64 hexagramas del Yi-King chino.







87. La primera cifra significativa
La cifras significativas de los precios del petróleo.



Director editorial
José Luis Córdoba
Asesor científico
Jaume Josa
Jefa de redacción
Ofelia Favarón
Secretaria de redacción
Natalia Franch
Traducción y Asesoramiento
Joan Pericay
Director de arte
Carlos Marino
Diseño gráfico
Dolors Cabecerán

La Recherche

Director general
Stéphane Khémis
Comité científico
Marc Augé • Claude Cohen-Tannoudji
Antoine Danchin
Jean-Gabriel Ganascia • Marc Jeannerod

PUBLICIDAD
Director de Publicidad
José Mª Barquín

Pérez Galdós 36
08012 Barcelona
Tel. 93 415 23 22 / Fax 93 238 07 30

EDITA
RBA Revistas S.A.
Pérez Galdós 36
08012 Barcelona (España)
Tel. 93 415 73 74* Fax 93 217 73 78

Presidente: Ricardo Rodrigo
Directores generales: Ana Rodrigo, Juan Manuel Rodrigo y José Luis Córdoba
Directora de Marketing Editorial: Mª Carmen Coronas
Directora Comercial: Ariadna Hernández
Director de Circulación: José Ortega
Director de Producción: Ricard Martínez
Jefe de Producción: Amadeu Granados

SUSCRIPCIONES
Servicio de Atención al Cliente
Pérez Galdós 36
08012 Barcelona
Nuevas suscripciones 902 392 391
Servicio de atención al cliente: 902 392 396
(de lunes a viernes de 9 a 14
y de 15 a 18 horas)
Fax 902 392 902

Director: Joan Muñoz
Distribución: Midesa. Tel. 91 662 10 00
Fotomecánica: FOINSA
Impresión-encuadernación: ROTOCAYFO, S.A-QUEBECOR.

El precio para Canarias, el mismo de la portada incluida sobretasa aérea.
ISSN 0211-3058



El hombre y el número

Desde el Antiguo Egipto, bastante antes que China, el número ha fascinado a los que piensan. Al principio eran los números enteros, la medida de las cantidades simples, de las distancias, de la edad y del tiempo que pasa, y el registro de las estrellas del cielo. Pero bien pronto hicieron su aparición los turbadores números irracionales.

En la Grecia del siglo VI, la escuela pitagórica basaba su reflexión sobre los números en un nuevo simbolismo que explicaba, como en China, todo el Universo, pasando por la música y la estructura de las constelaciones. Un siglo más tarde, con Sócrates, enemigo irreductible del pensamiento mágico, nacía la corriente racional propiamente dicha, primer milagro que conduciría, dos siglos más tarde, al desarrollo de la Escuela de Alejandría, con Euclides, Arquímedes y hasta Diofanto.

Compás de espera. Durante siete siglos, las matemáticas occidentales permanecen estancadas. Los sabios siguen pegados al engrudo de las cifras romanas y de la exégesis.

Un segundo milagro tendría lugar a finales del siglo XI, en el sur de Europa, los jerarcas de la Iglesia traducen los tratados árabes inspirados en la tradición india. Siguen luego Fibonacci y Pacioli; más tarde, la explosión conceptual del Renacimiento, seguida de Kepler, Descartes, Fermat, Leibnitz, Newton...

A pesar de los reproches de Descartes, e incluso en él mismo, el pensamiento racional y el pensamiento mágico permanecen mezclados, pero el primero progresa tan rápidamente, su deseo de vivir es tan intenso, su eficacia es tan clamorosa, que, en el siglo XVIII, termina por invadir todo el ámbito de la reflexión.

En el siglo siguiente, Galois descubre (¿o inventa?) el concepto de grupo y Cantor la teoría de los conjuntos. El zoo de los números y de las teorías se enriquece de manera casi exponencial.

Otro siglo y he aquí la informática, que trastorna las condiciones y las posibilidades del cálculo.

Los ámbitos de exploración matemática continúan diversificándose y haciéndose más complejos, hasta el punto de que ya no basta con ser un matemático profesional para comprender lo que se está haciendo en los múltiples sectores de su propia disciplina.

El número, además, es explotado por las administraciones y las empresas para afirmar su poder, y, en el mismo seno de la comunidad de los matemáticos, continúa suscitando interrogantes, los mismos que ya se formulaban los filósofos griegos.

¿Podría ser que el número fuese una realidad en cierta manera superior, que no sólo preexistiría a la escritura, sino también al hombre e incluso a los elementos?

¿Qué es un número?

Christian Houzel

Especialista en espacios analíticos e historiador de las matemáticas, Christian Houzel nos relata las múltiples metamorfosis de la noción de número, desde los primeros sistemas de contar de los sumerios hasta las últimas especies numéricas inventadas por los matemáticos. La vieja pregunta es: ¿cabe definir un número?

Mundo Científico: ¿Es posible identificar históricamente el momento en que emerge la noción abstracta de número?

Christian Houzel: Si, tenemos esta suerte. En la antigua Mesopotamia, los sumerios disponían de notaciones diferentes según que se tratara de objetos que se podían contar uno a uno, como las cabezas de ganado, de medidas de volumen para líquidos, para grano, y así sucesivamente. Y luego, hacia finales del III milenio, los escribas sumerios inventaron un sistema destinado a aplicarse a cualquier cosa susceptible de ser contada. Aquí es donde vemos formarse el concepto de número abstracto en la notación escrita. Los sumerios unificaron todas las notaciones existentes en base sesenta.

¿Por qué sesenta?

Este número viene de uno de los sistemas metrológicos anteriores, en el que los multiplicadores, para pasar de una unidad al grupo de la escala superior, eran alternativamente seis y diez. Se procedió pues a una reagrupación. Los babilonios heredaron esta notación en base sesenta, que nosotros hemos heredado también en nuestra manera de contar las horas, los minutos y los segundos.

¿Conocían los babilonios los números racionales?

No, como tampoco los egipcios. Es verdad que ambos disponían de ciertas formas de fracciones. Pero fracciones de numerador 1. Con su sistema de base sesenta, los babilonios escribían de la misma manera lo que hoy son las cifras que van después de la coma. El clavo vertical que designaba la unidad lo mismo podía designar sesenta unidades que una sesentava parte.

Las fracciones egipcias son lo que a veces se conoce como «cuantésimos»: ante un problema como el de repartir una ración entre varios obreros, cuando el resultado no es exacto, las partes fraccionarias se expresaban como combinaciones de fracciones de tipo un cuarto, un octavo, un catorceavo, etc.

> Sólo en el siglo XIX se fundó la noción de número real de un modo puramente aritmético, prescindiendo de la geometría

Hubo que esperar pues a la Grecia antigua para que aparecieran los racionales...

En realidad tampoco se puede decir esto, porque para los matemáticos griegos los únicos números identificados como tales eran los enteros. Fue la escuela pitagórica, no sabemos exactamente en qué época, lo más tarde en el siglo V antes de nuestra era, la que descubrió realmente la irracionalidad de ciertas razones de magnitudes. En particular, la diagonal del cuadrado guarda una relación irracional con el lado, lo cual significa que este cociente no puede ser igual al de dos números enteros. Pero ante el problema planteado los griegos no inventaron un nuevo tipo de números sino una teoría de las proporciones entre magnitudes geométricas completamente independiente de los números. Es lo que se encuentra en el libro V de los Elementos de Euclides: una teoría que permite manejar proporciones entre magnitudes geométricas. Pero estas magnitudes nunca se consideran como medidas por números. Por tanto, no se puede decir que $\sqrt{2}$ existe en la matemática griega. Lo que existe son magnitudes geométricas y cocientes entre magnitudes geométricas, que no necesariamente son racionales.

¿Tampoco π era un número?

No. En el libro XII de los Elementos de Euclides hay una proposición que dice que el área de un círculo es proporcional al cuadrado del radio del círculo. Se puede pues concebir, en el marco de la geometría griega, el cociente entre el área del círculo y el cuadrado del radio. Es lo que llamaríamos π. Pero este tipo de razón no se contempla como un objeto matemático. Se trata de relaciones entre magnitudes. Aunque Arquímedes calculó una aproximación de π, todavía no había número π en la matemática griega.

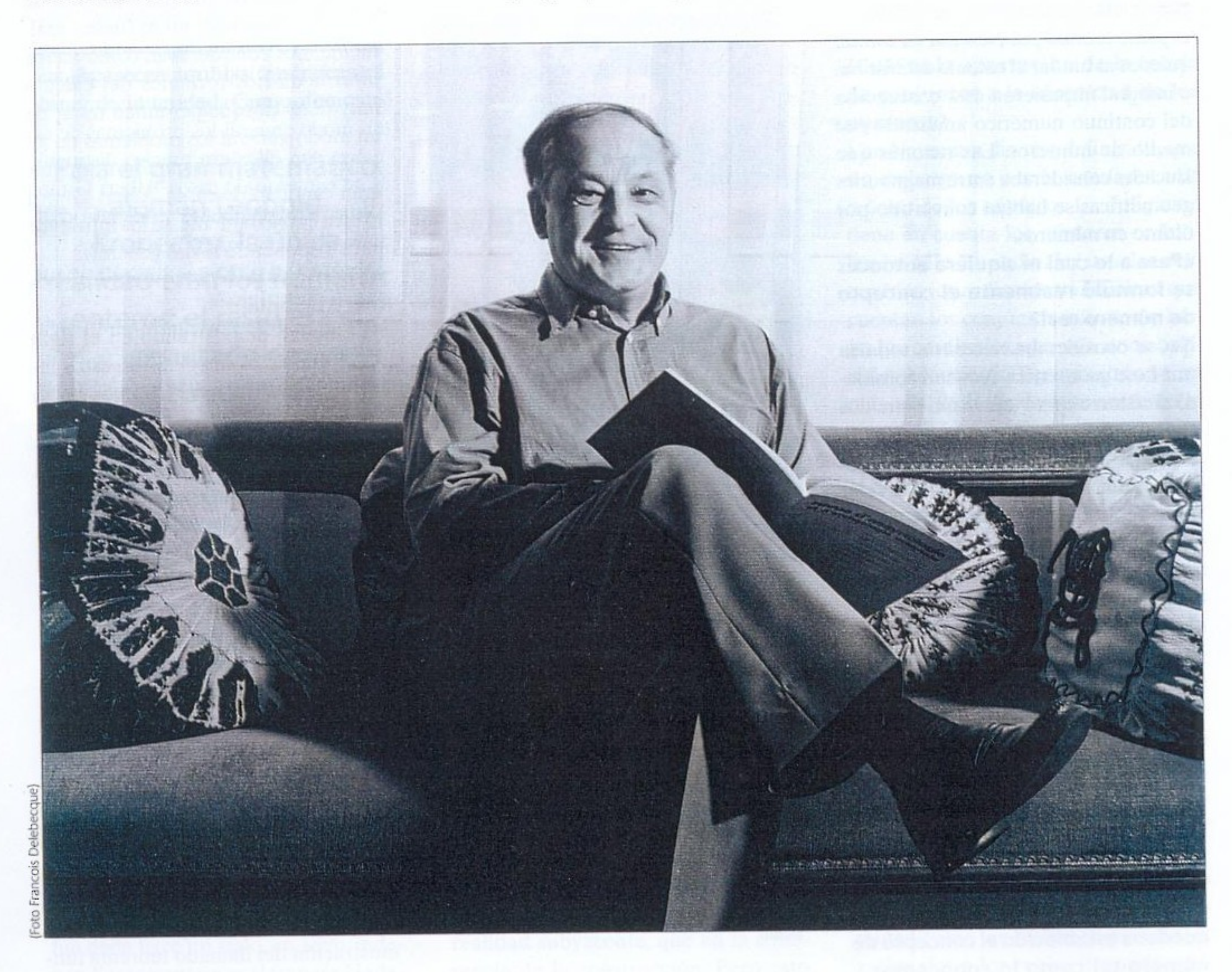
¿Pero fue en la Grecia clásica donde surgió la primera reflexión sobre la identidad del número?

Indudablemente, pero no sabemos cuándo ni cómo. El primer texto matemático que poseemos, estos famosos Elementos de Euclides, datan del 300 antes de nuestra era, una fecha ya tardía respecto a la Grecia clásica. No conocemos con exactitud la génesis de las concepciones que allí se expresan. Euclides define un número como una multiplicidad de unidades, siendo la unidad un concepto primitivo no definido, que ya está ahí. Y en virtud de la

teoría de las proporciones, la geometría goza de una cierta preeminencia sobre el resto de las matemáticas. Cuando Euclides trata la teoría de números, en sus libros VII, VIII y XIX, representa los números por medio de segmentos, lo que le permite razonar sobre números no especificados, abstractos. Ni siquiera dice si se trata de 2 o de 3, lo que dice es: considero un número y lo designo por medio de un segmento. Y a estos segmentos les aplica razonamientos geométricos. Es pues la geometría la que sirve a la teoría de números.

árabes, que introdujeron el concepto fundamental de ecuación. Ecuación en el sentido de expresión de un problema.

En la intención de sus fundadores, esencialmente al Khāwārizmi, a comienzos del siglo IX, un cierto número de problemas matemáticos podían expresarse en forma canónica: lo que hoy llamaríamos ecuación de segundo grado. Estaba la incógnita, que al Khā wārizmi llamaba *la cosa*, y luego el cuadrado de la incógnita, que llamaba, en árabe, *el bien, la riqueza* (*census* en latín) —porque este tipo de consideraciones las aplicaba sobre todo a problemas de reparto de herencia. Una ecuación era pues una relación que combinaba la incógnita, su cuadrado y un número. Se trataba de hallar la incógnita. Pero esta álgebra se utilizó también, desde el principio, para resolver problemas de tipo numérico en un contexto geométrico, para resolver problemas en los que la incógnita era una magnitud geométrica. Después de lo cual los matemáticos árabes desarrollaron un cálculo de polinomios, introdujeron otras potencias, el cubo, la cuarta potencia, e incluso potencias negativas. Pero dependiendo del problema la incógnita podía ser una magnitud geométrica o un número. Esto indujo una reacción en sentido contrario: los árabes se dieron cuenta de que podían tratar de un modo calculatorio las propias magnitudes geométricas. Reinterpretaron a la manera aritmética el libro X de los Elementos de Euclides, que presentaba toda una teoría sobre la clasificación de las magnitudes irracionales que se encuentran en las construcciones geométricas. Luego, a partir del siglo XII, los algebristas árabes efectuaron cálculos aproximados en notación decimal con cifras detrás de la coma, unos cálculos muy elaborados con raíces de orden cualquiera.



(Foto Francois Delebecque)

Cuando Arquímedes calcula su aproximación de π, ¿sigue tratándose, para él, de expresar una razón geométrica, no numérica? Pero entonces, ¿cuándo se opera la fusión entre las dos concepciones?
Se hace muy tarde, hacia fines del siglo XVIII, con gente como Newton y Leibniz, al término de una larga maduración. Tiene su origen en la práctica algebraica. El álgebra fue fundada en el siglo IX por los matemáticos

Pero todavía en el siglo XVII, cuando Pascal escribió: «La geometría no puede definir los números, el movimiento ni el espacio», ¿consideraba que por un lado había los

números y por el otro lo que llamaba «espacio», es decir, lo que llamamos hoy «geometría»?
Indudablemente. La matemática europea vivió hasta el siglo XVII sobre esta teoría de las proporciones euclídeas. Por lo demás, sólo en este época se adoptaron los números decimales con cifras después de la coma, que pasaron a ser de uso corriente en los cálculos astronómicos. También a principios del XVII se inventaron los logaritmos, lo cual indicaba ya una especie de concepción de un continuo numérico, pero aparecieron por razones prácticas antes de ser aceptados en el plano teórico por Newton y Leibniz, quienes al fundar el cálculo diferencial e integral impusieron esta concepción del continuo numérico analizado por medio de números. Las razones que Euclides consideraba entre magnitudes geométricas se habían convertido por último en números.

¿Pese a lo cual ni siquiera entonces se formuló realmente el concepto de número real?
No, se consideraba necesaria todavía una base geométrica. Newton consideró al número como una razón entre dos cantidades homogéneas de misma naturaleza. Concebía las razones geométricas como números, pero esta concepción se basaba todavía en la geometría. Muy poco a poco, ya en el siglo XIX, se fue advirtiendo que se podían fundar los números reales en consideraciones puramente aritméticas, prescindiendo de la geometría. El primero en tener la idea fue Bolzano, un filósofo matemático checo, pero su tentativa no culminó. El primero en realizar las construcciones aritméticas de los números reales fue Dedekind en 1858, al que siguieron Weierstrass y luego Cantor en 1872.

¿Se puede considerar que a partir de entonces, a fines del siglo XIX, quedaba establecido el concepto de número tal como lo conocemos y practicamos hoy en día?
Sí, pero con matices ligados a la naturaleza atribuida a los distintos tipos de números. Había en particular los números imaginarios y complejos, que era necesario usar para estudiar los enteros. Los números imaginarios aparecen por primera vez en un libro de álgebra de Girolamo Cardano, un matemático del norte de Italia, en el siglo XVI. Pero no les sacó partido. El primer texto en el que vemos realmente en acción a estos números es un poco posterior. Es de otro italiano, Bombelli, en 1572. Se trata simplemente de raíces cuadradas de números negativos. Hay una regla de signos conocida desde hace mucho tiempo que dice que si se multiplica más por más, o menos por menos, se obtiene siempre más. Por tanto, un cuadrado de un número en sentido ordinario es necesariamente positivo. Bombelli se enmarcaba en el desarrollo aritmético de la teoría de los irracionales, que tiene su origen en el álgebra árabe. Bombelli le da mayor extensión; dice haber descubierto un nuevo tipo de irracional, le da unas reglas de cálculo y demuestra que dichos números son útiles para estudiar la ecuación de tercer grado. Era una especia de cálculo formal, que no era en absoluto interpretable geométricamente. Más tarde, en el siglo XVII, Albert Girard en 1629 y luego René Descartes enunciaron que una ecuación algebraica tiene tantas raíces, o soluciones, como grados. Una ecuación de segundo grado tiene dos raíces, una de tercer grado, tres, etc. En su *Geometría* de 1637, apéndice del famoso *Discurso del método*, Descartes, como Girard antes qué él, explicó que el número de raíces era igual al grado, pero que estas raíces no siempre eran reales. Fue allí donde introdujo el término «imaginario» (Girard decía «imposible»). Imaginario, es decir, que cabe imaginar un número de soluciones igual al grado. Cabe imaginarlos, es decir, representarlos por letras, ya que se está en el marco de un álgebra literal, y manipularlos como si fueran números, aunque a los ojos de Girard y Descartes no lo fuesen realmente. Los consideraban como meros intermediarios formales del cálculo muy útiles porque permitían tratar de un modo general los problemas de álgebra. Sin ellos era preciso distinguir gran cantidad de casos. Albert Girard lo dice explícitamente: se los introduce para disponer de reglas generales.



Fragmento de un Códice del siglo XIV que representa a las siete Artes Liberales (Biblioteca Ambrosiana de Milán).

¿Y cómo se pasó a la noción de número complejo?
Se empezó demostrando en el siglo XVIII que estos imaginarios, estos intermediarios formales —pues no se sabía muy bien qué eran— eran utilizables para el cálculo integral.
Pero entonces ya no se podía decir simplemente que eran meros objetos formales, puesto que había que escribir, por ejemplo, el logaritmo de un imaginario. Había que tratar de interpretar qué podían ser. En la práctica, el número imaginario intervenía siempre en un par de números reales. Es este par lo que llamamos número complejo. Un número complejo comprende una cantidad real, digamos p, y otra cantidad real, digamos q, multiplicada por √-1. La única intervención de la imposibilidad es la introducción de √-1. Fueron los números complejos los que permitieron a d'Alembert, en 1746, emprender la demostración del llamado teorema fundamental del álgebra. D'Alembert precisó el enunciado de Descartes según el cual una ecuación tiene tantas raíces como grados demostrando que las raíces imaginarias introducidas por el filósofo-matemático son todas de la forma p + q …–1. La demostración completa se debe a Gauss (1799).

¿Por qué los números complejos son indispensables para estudiar los números enteros?
El caso típico es el del último teorema de Fermat, un contemporáneo de Pas-

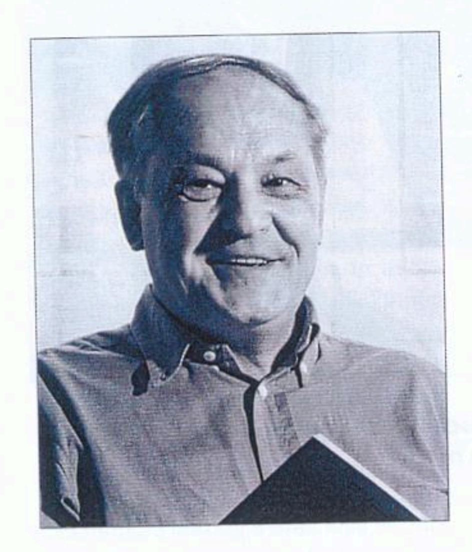
cal, demostrado hace sólo cinco años (véase el artículo de Catherine Goldstein en este número). El teorema dice que cuando n es un entero mayor que 2 no hay enteros positivos a, b y c que verifiquen la ecuación $a^n + b^n = c^n$. El caso más simple es que una suma de dos cubos no puede ser un cubo. Ahora bien, este caso ya exige la intervención de números complejos. La suma de dos cubos puede descomponerse en factores. $x^3 + y^3$ es divisible por $x + y$. El cociente es $x^2 - xy + y^2$. Esta fórmula puede volver a factorizarse, pero hay que usar números complejos, pues se obtiene una ecuación de segundo grado que no tiene raíces reales. Aparecen aquí las tres raíces cúbicas de la unidad. Como demostró

Para el gran matemático alemán Leopold Kronecker, la única realidad eran los números enteros; el resto era el resultado de construcciones

Euler en el siglo XVIII, también es necesario introducir números complejos para estudiar la sucesión de los números primos. Todavía hoy, uno de los problemas centrales de las matemáticas consiste en saber cómo están distribuidos los números primos en la sucesión de los enteros. Estos números aparecen de manera inesperada. Sabemos que son cada vez menos numerosos a medida que se avanza en la sucesión de los enteros, pero no logramos determinar la regla que preside su aparición. El primer teorema que dio una información sobre el modo como estos números se van haciendo cada vez menos numerosos fue dada hace un siglo, en 1896, independientemente por el francés Hadamard y el belga La Vallée-Poussin. Para obtenerlo, hubo que utilizar las propiedades analíticas de una cierta función, la llamada función zeta de Riemann, que ya había sido introducida por Euler en el siglo XVIII, y que recurre a números complejos.

No se puede decir, por tanto, que un número imaginario tenga menos realidad que un número entero...

No. Pero de todos modos subsisten puntos de vista diferentes sobre el grado de realidad de los distintos tipos de números. Estas divergencias vienen de lejos. La matemática clásica, de la Antigüedad a principios del siglo XIX, tenía una base ontológica, que era situada en distintos lugares según los matemáticos. Para Euclides, los objetos matemáticos eran probablemente ideas platónicas. En la filosofía de Platón, las ideas son las verdaderas realidades y lo que nos rodea no es más que un reflejo de las verdaderas realidades. Para los filósofos de la Ilustración como d'Alembert, los objetos matemáticos eran considerados más bien como poseedores de una base empírica, como abstracciones sacadas del mundo sensible, del mundo físico que nos rodea. Pero en cual-



quier caso había siempre una base ontológica. Ésta tendió a disolverse cuando en el siglo pasado se descubrió la existencia de geometrías no euclídeas. Hay varias geometrías posibles y ningún medio para decidir si una es más real que otra. Desde aquella época, por tanto, para la mayoría de los matemáticos el lugar de la verdad de las matemáticas está menos en la idea de una base ontológica, de una realidad subyacente, que en la coherencia de la construcción. Pero esto depende de los autores. Para Kronecker, uno de los grandes teóricos de los números en el siglo XIX, la única realidad eran los números enteros. Todo el resto era resultado de construcciones realizadas por los matemáticos. Incluso hoy, para un matemático como René Thom, la única realidad es el continuo, y por tanto los números reales. Para él los números enteros vienen en segundo lugar, pues están sacados del continuo. Es el punto de vista inverso al de Kronecker. Estos posicionamientos tienen siempre una carga ideológica ligada a la trayectoria del matemático. Alexander Grothendieck, geómetra como Thom pero de otra tendencia, fundó la geometría algebraica sobre bases completamente distintas, donde los números reales carecen de verdadero lugar. No lo digo para minimizar el punto de vista de Thom, que es muy interesante. Es verdad que no se pueden imaginar los números enteros haciendo abstracción del continuo. Cuando se cuenta se cuenta en el tiempo y hay necesariamente un continuo en alguna parte...

¿Cuál es la aportación de la teoría de conjuntos a la identidad de los números?

Von Neumann dio una definición, que sin embargo no puede satisfacer a un teórico de los números. La definición tiene en cuenta los números ordinales, no sólo los ordinales usuales sino también los ordinales transfinitos, que cuentan los conjuntos infinitos, y luego los cardinales, que son unos ordinales especiales. Pero cuando se va más allá de lo finito y se entra en lo transfinito los ordinales son muy especiales, hay muchísimos que tienen el mismo cardinal, la situación es completamente diferente. En cualquier caso, se trata de números en el sentido discreto del término; sirven para contar. Los números reales no tienen cabida aquí. Los números de la teoría de números tampoco son exactamente los mismos, pues trabajamos con números racionales, irracionales, irracionales complejos... y aún otros irracionales recientemente introducidos, como los llamados números p-ádicos. Los números p-ádicos son unos nuevos irracionales inventados para completar los racionales usuales. Tienen en cuenta no propiedades de proximidad como los reales (se pasa de los racionales a los reales llenando agujeros) sino de divisibilidad por los números primos. Es otra manera de llenar agujeros, una magnífica idea (véase el artículo de Daniel Barsky y Gilles Christol en este número).

¿Se enriquecerá todavía más el zoo de las especies numéricas?

Sin duda alguna. Hay unos nuevos números, muy interesantes, llamados surreales, que han sido inventados por el inglés John Conway. Conway partió de la idea que tuvo Dedekind para



La «Perla filosófica» de Gregor Reisch (Freiburg, 1503). La aritmética alegórica arbitra en favor del cálculo por cifras.

construir los números reales. Para Dedekind, un número real es una cortadura. Un número real corta el conjunto de todos los números racionales en dos pedazos, uno a la izquierda y el otro a la derecha, de modo que cualquier elemento de la izquierda es inferior a cualquiera de la derecha. Para Conway, esta operación de cortadura es inútil. Un número en el sentido de Conway es un par de conjuntos tales que cada elemento del primero sea más pequeño que cada elemento del segundo. Conway parte del conjunto vacío, luego explica qué quiere decir con mayor y menor, y poco a poco va construyendo todos los enteros, los racionales, y luego los números infinitamente pequeños e infinitamente grandes. Es una especie de curiosidad marginal pero muy bonita.

Como los números primos, algunos nuevos tipos de números parecen imponerse a los matemáticos con una fuerza comparable al menos a la de los objetos del mundo físico.

Claro. Lo que diría de la realidad es que se está ante algo real cuando este algo resiste. El matemático, como el físico, se ve confrontado al descubrimiento de nuevos cuerpos celestes, nuevas galaxias... Las cosas resisten. No hacemos lo que queremos. Los números primos están ahí en la sucesión de los números enteros, no hacemos otra cosa que descubrirlos. Pero no podemos decretar que tal o cual número será primo. Los números son lo que son y es un duro trabajo llegar a detectarlos, a conocer sus propiedades. Se está ante una realidad ciertamente tan real como como el mundo que nos rodea y en cierto modo más real, porque se accede a ella con mayor claridad. Cuando se ha obtenido un resultado matemático, este resultado es absolutamente innegable y completamente comprensible. Mientras que cuando se ha obtenido un resultado sobre tal galaxia, sobre tal partícula elemental o tal hecho biológico, el resultado sigue estando sometido a discusión, a interpretación. No digo que en matemáticas la claridad sea siempre total. No hay que creer que porque se obtuvo una construcción aritmética de los números reales en el XIX los problemas filosóficos del continuo estén todos resueltos. Por ejemplo, en los años 1960, el lógico inglés Abraham Robinson inventó el llamado análisis no estándar, que es una especie de extensión de los números reales en la que además de los números usuales se introducen los números infinitamente pequeños y los infinitamente grandes. Como se ve, todavía podemos enriquecer el continuo, hallar cosas nuevas.

¿Está usted diciendo que la sucesión de los números primos tiene una realidad más estable que la realidad material que nos rodea?

Sí, sin duda alguna, pues lo que conocemos lo conocemos de un modo totalmente cierto. Mientras que en las ciencias experimentales hay inevitablemente un elemento de inestabilidad.

¿Es para usted una realidad distinta del mundo material?

Es una realidad que no pertenece al mundo sensible. No es el mundo material. No hay ninguna relación con los cinco sentidos.

¿Hay que concluir que no sabemos muy bien qué son los números?

Claro. No hay que engañarse al respecto. Actualmente, todavía se presentan los fundamentos de las matemáticas sobre la base de la teoría de conjuntos. Pero es la fase actual de las matemáticas. Después del fracaso de varias tentativas, sabemos perfectamente que no podemos construir completamente los fundamentos de las matemáticas. Estamos obligados a tener una actitud un poco más pragmática: hacer matemáticas sin preocuparse excesivamente por los fundamentos. Sabemos perfectamente que los fundamentos últimos de las matemáticas son inaccesibles. Por lo que, en última instancia, no sabemos qué es un número.

Declaraciones recogidas por **OLIVIER POSTEL-VINAY**

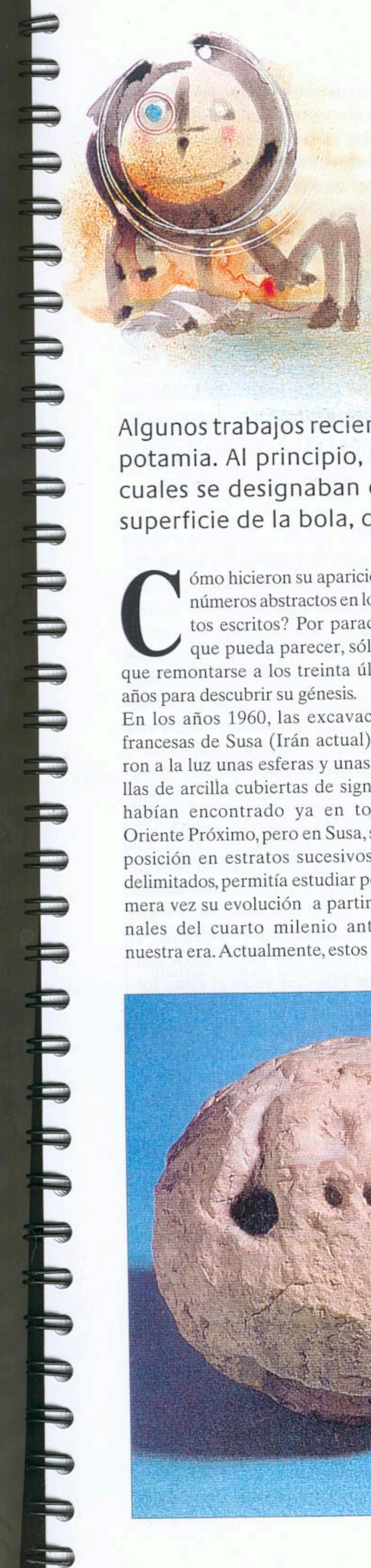
Para Más Información:

J. Ritter, «Metrology and the prehistory of fractions», *Histoire des fractions d'histoire*, Birkhauser, Basilea, 1992, p. 19-34.

R. Dedekind, *Les Nombres, que sont-ils et à quoi servent-ils?* traducido el alemán por J. Milner y H. Sinaceur, Navarin, Pans, 1979.

R. Rashed, *Entre arithmétique et algèbre*, Les Belles Lettres, Paris, 1984.

B. Torrecillas Jover, *Fermat. El mago de los números*, Ed. Nivola, Madrid, 1999.



El nacimiento del número

Catherine Goldstein

Algunos trabajos recientes permiten reconstruir la génesis del número escrito en Mesopotamia. Al principio, eran unas bolas huecas de arcilla que contenían fichas con las cuales se designaban cantidades de bienes. Luego, estas fichas se imprimieron en la superficie de la bola, que más tarde se aplanó para dar lugar a una tablilla.

Cómo hicieron su aparición los números abstractos en los textos escritos? Por paradójico que pueda parecer, sólo hay que remontarse a los treinta últimos años para descubrir su génesis.

En los años 1960, las excavaciones francesas de Susa (Irán actual) sacaron a la luz unas esferas y unas tablillas de arcilla cubiertas de signos. Se habían encontrado ya en todo el Oriente Próximo, pero en Susa, su disposición en estratos sucesivos, bien delimitados, permitía estudiar por primera vez su evolución a partir de finales del cuarto milenio antes de nuestra era. Actualmente, estos signos se interpretan como precursores de la escritura, gracias a las investigaciones fundamentales del arqueólogo Pierre Amiet (por entones conservador-jefe de antigüedades orientales del Museo del Louvre). Algunas hipótesis, contestadas aunque estimulantes, de la arqueóloga norteamericana Denise Schmandt-Besserat sobre la naturaleza exacta de estos signos, y el renacimiento de los estudios sobre las matemáticas mesopotámicas en los años 1980, gracias a los esfuerzos conjuntos de asiriólogos y de historiadores de las ciencias, permiten ahora efectuar un análisis convincente de la aparición progresiva del número «abstracto» en los textos escritos. Al menos en lo que se refiere a las civilizaciones (Sumer, Elam) que florecieron más de tres mil

Figura 1. La fotografía en blanco y negro es una tomografía con rayos X de una «burbuja» de arcilla (Uruk, mediados del IV milenio). Este procedimiento permite visualizar las fichas contenidas en la bola sin tener que romperlas. La fotografía en color muestra una bola, originaria de Susa, datada el 3300 a.C., y las fichas que contenía (Museo del Louvre). Estos objetos servían como registro de contabilidad. (Foto blanco y negro: P. Damerow y H.P. Meinzer; foto color ©RMN).

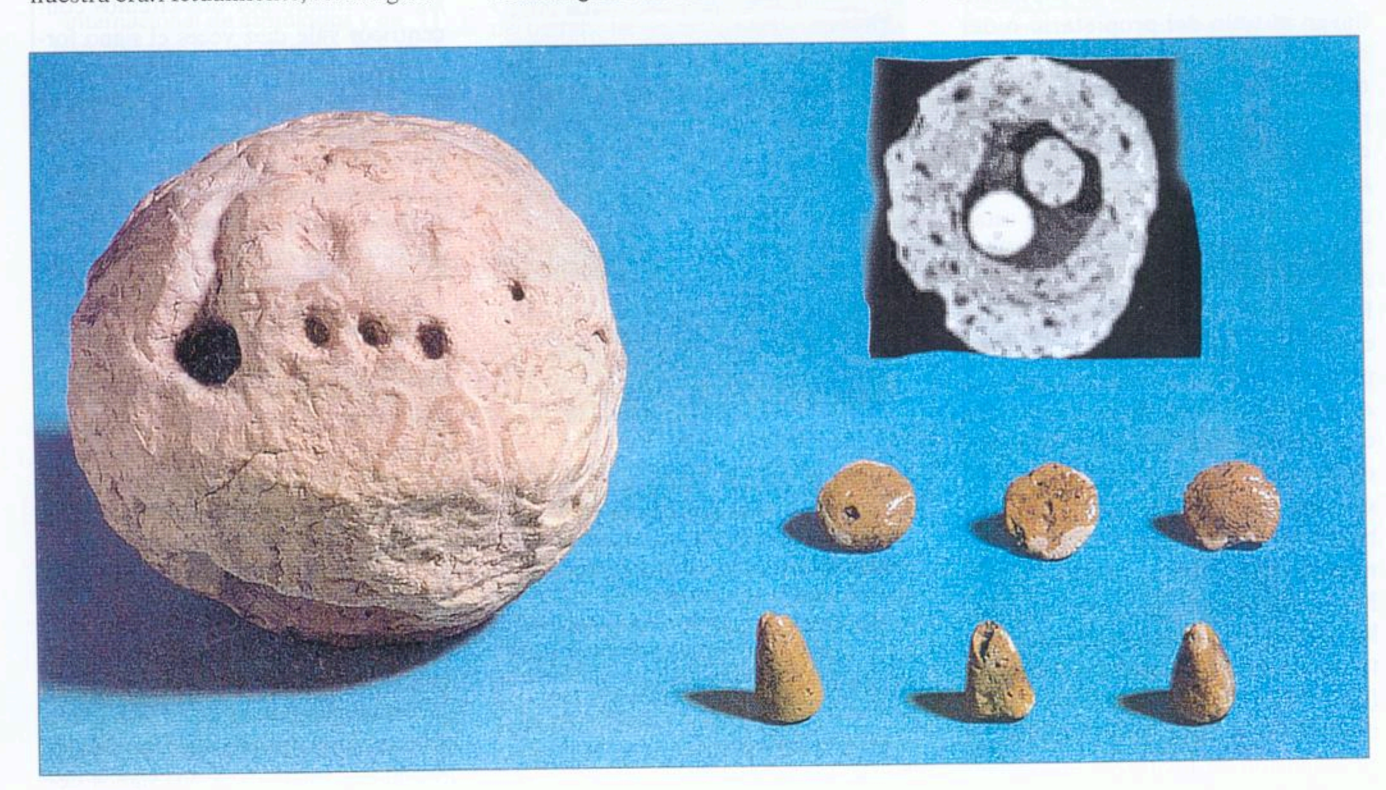




Figura 2. Tablilla de contabilidad de raciones (Uruk, hacia 3000 a.C.). **Las fichas se han sustituido por marcas impresas en la arcilla mediante un cálamo de caña.** (Foto P. Damerow).

años antes de nuestra era en los territorios de Irán e Irak actuales y que utilizaron para la escritura un soporte sorprendentemente duradero: la arcilla.

Al principio se empleaban unas «burbujas», en realidad unas bolas huecas, y unas fichas de arcilla. En el período más antiguo, las bolas contenían unas fichas pequeñas (fig.1), las cuales, con sus tamaños y formas variadas, representaban diversas cantidades de bienes (corderos, medidas de aceite o de trigo). Sin duda, las bolas eran una especie de registros primitivos de contabilidad; en su superficie llevan el sello del propietario o del controlador, o bien el de las partes contratantes. Se supone que se rompían para verificar su contenido. En una etapa posterior, seguramente para permitir controles intermedios sin tener que romper la esfera, las fichas estaban hundidas en la arcilla de la bola antes de ser encerradas en el interior. Por tanto, en esta fase hay unos signos exteriores a modo de duplicados de pequeños objetos representativos.
Aproximadamente, hacia finales del cuarto milenio antes de nuestra era, los objetos-ficha desaparecen. Quedan tan sólo sus marcas en la superficie de la bola. Es estonces cuando ésta se aplana para convertirse en tablilla (fig. 2). Muy rápidamente, para grabar todas las marcas exteriores se utiliza un instrumento único y especial, el cálamo (de caña). La variación de los signos necesarios se obtiene mediante la combinación de los trazos que dejan los dos extremos del cálamo, de tamaño desigual: hundidos verticalmente, se obtienen círculos; transversalmente, dan entalladuras. Por fin, en una última etapa, el extremo del cálamo indica las diferentes cantidades de bienes, mientras que un dibujo más elaborado completa la información, precisando la naturaleza de las mercancías tratadas (ganado, trigo). Se consuma entonces una separación entre el signo escrito



Figura 3. Hacia 2000 a.C., se adopta en Mesopotamia un sistema unificado de escritura y de contabilidad. Aparece la escritura cuneiforme, y el sistema de numeración de base 60 se convierte en el dominante. Esta tablilla (abajo) es una tabla de multiplicar por 25, procedente de Susa y conservada en el Museo del Louvre.
(Foto Ch. Larrieu/La Licorne).

cuantitativo y el signo escrito cualitativo: uno y otro continúan independientemente su evolución rápida hacia ... las matemáticas por un lado y la literatura por otro. Hay que destacar que, en esta etapa precoz, la escritura no representa (todavía) la lengua hablada, ya que prácticamente carece de estructura sintáctica.

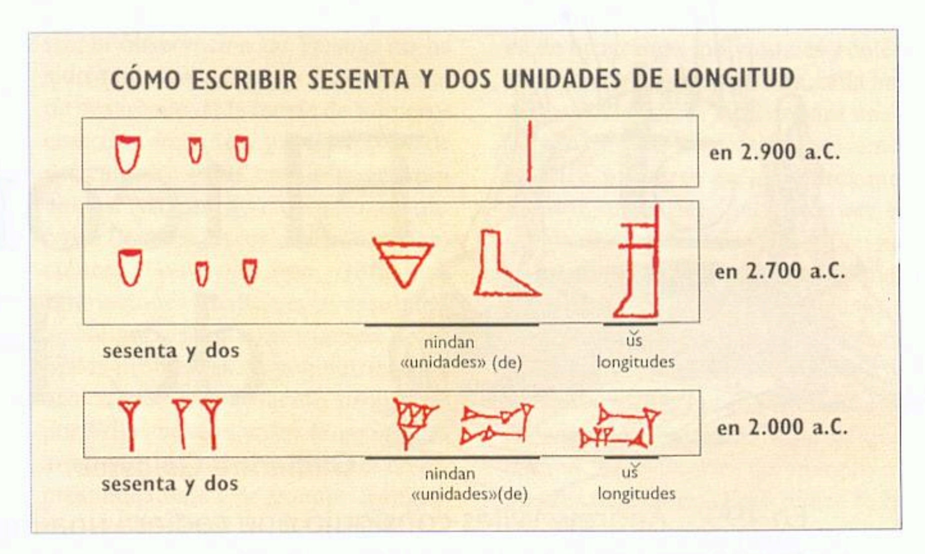
Hacia 3000 a.C., los signos numéricos se organizan en una docena de sistemas diferentes. Uno designa cantidades discretas, otro unidades de superficie...

Pero, ¿qué puede decirse de los signos cuantitativos? Durante el período llamado arcaico (3200-2800 a.C.), los signos numéricos se organizan en una docena de sistemas metrológicos diferentes. Sus valores y sus relaciones sólo han empezado a comprenderse gracias al trabajo pionero del sueco Jöran Friberg y de los alemanes Peter Damerow y Bob Englund en los años 1980. Se ha observado que hay un sistema (llamado sistema S) para las cantidades discretas, por ejemplo, corderos; otro (llamado sistema G) para la medida de las superficies de terreno, etc. El mismo signo puede representar números diferentes de unidades según los distintos sistemas. Por ejemplo, el signo formado por dos círculos concéntricos vale diez veces el signo formado por un círculo grande, si se trata de ovejas; pero si lo que se desea es medir la superficie de un campo, es, por el contrario, el círculo grande el que vale seis veces los dos círculos concéntricos. Por tanto, los signos numéricos no tienen ningún valor intrínseco, sino que dependen del sistema metrológico en el que se insertan.

En el período llamado protodinástico (2800-2350 a.C.), se introducen reformas en el sistema de escritura, a causa del tamaño creciente de los intercambios entre las ciudades-estado de Mesopotamia y de Elam. El número de sistemas metrológicos utilizados disminuye, la escritura se desarrolla y permite la reproducción del lenguaje hablado. Para resolver ciertas ambigüedades residuales de los antiguos sistemas (por ejemplo, el hecho de que los valores relativos de los signos de-

pendan del sistema considerado), el número de la unidad correspondiente se escribe de una manera explícita. De este momento datan los primeros textos propiamente matemáticos de que disponemos: son unas tablillas y unos ejercicios escolares destinados a la formación profesional del futuro escriba.

Durante el período siguiente (hasta principios del segundo milenio), se crean imperios centralizados que ponen en funcionamiento, a mayor escala, un sistema unificado de escritura y de contabilidad. La escritura se simplifica para acelerar el trazado : es el momento en que aparece la escritura cuneiforme, bien conocida (foto 3). Pero, a la vez, surgen nuevos riesgos de ambigüedad, ya que ciertos signos (grandes y pequeñas entalladuras, por ejemplo) tienden a confundirse. La solución que se adopta a finales de este período es aceptar como dominante un sistema numérico único, inspirado en el el sistema S. Más exactamente, sirve para escribir todos los números durante los cálculos; la conversión a otra unidad y la mención de la unidad empleada se añaden por separado al final. Este sistema es el de posición en base sesenta (véase esquema). Un número que se escribiera 421 en este sistema —naturalmente, los signos utilizados en Mesopotamia son distintos de nuestras cifras arábigas— equivaldría a 4 x 3600 + 2 x 60 + 1, es decir, catorce mil quinientos veintiuno. Nosotros aún conservamos restos de este tipo de sistema en la medida del tiempo en horas, minutos y segundos.



Este esquema muestra cómo evolucionó la escritura de los números en la antigua Mesopotamia, desde las entalladuras hechas con el cálamo hasta la escritura cuneiforme.

A partir de este momento, quedan establecidos los números abstractos, cuya escritura y manejo ya no dependen de los objetos que designan. Pero para ello, le ha sido preciso liberarse poco a poco de otros signos (nombres de bienes; luego, unidades metrológicas), siempre bajo la presión de unos condicionantes muy concretos; algunos de ellos estaban relacionados con factores exteriores, como la racionalización de la metrología, mientras que otros dependían de anteriores opciones escogidas dentro de los sistemas de numeración establecidos. Sería importantísimo poder comparar esta evolución con la de otras civilizaciones. Desafortunadamente, de muchas de ellas, como la de China o de la India, no disponemos de textos suficientemente antiguos para analizar estas etapas precoces. Es de esperar que el trabajo que se está realizando en otros casos (especialmente sobre Egipto), proporcione elementos de comparación para determinar los factores cruciales en la aparición del concepto de número.*

CATHERINE GOLDSTEIN está a cargo de las investigaciones del CNRS y trabaja en el laboratorio de aritmética y de geometría algebraica (URA D752) en la Universidad Paris-Sud. Actualmente se dedica a la historia de la teoría de los números en el Occidente moderno.

El autor agradece a Peter Damerow, Hans Nissen y Jim Ritter todas las informaciones y documentos que le han facilitado para la redacción de este artículo.

EL «GRUPO DE BERLÍN» Y EL INICIO DE LOS NÚMEROS

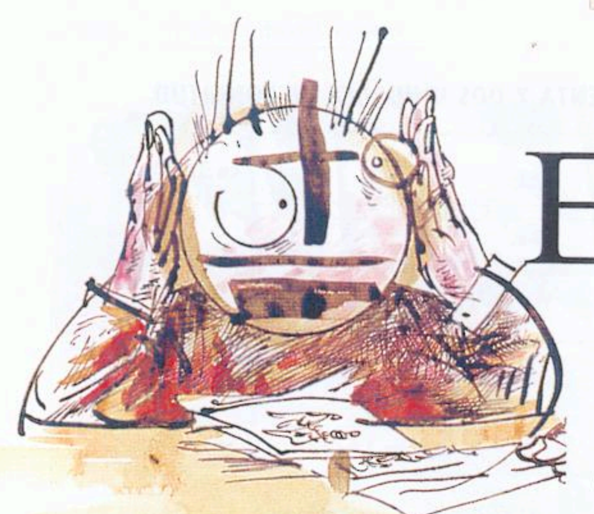
Desde hace algunos años, nuestra comprensión del inicio de los números se ha renovado gracias a los trabajos de un grupo internacional de asiriólogos y de historiadores de las ciencias que se reúnen regularmente en el marco del «Berlin Workshop».

Forman parte de este grupo Marvin Powell (Universidad de Illinois, Estados Unidos), Jöran Freiberg (Universidad de Göteborg, Suecia), Jens Høyrup (Universidad de Roskilde, Dinamarca), Peter Damerow (Instituto Max Planck, de Berlín), Bob Englund y Hans Nissen (Universidad Libre de Berlín), Jim Ritter (Universidad de París 8), etc.

Sus investigaciones han demostrado el interés que tiene estudiar de cerca los sistemas metrológicos, con frecuencia desdeñados, para comprender la evolución de las ideas aritméticas. Este grupo de investigadores pone en práctica medios técnicos muy elaborados (escanerización, tratamiento informático, etc.) para recoger y analizar centenares de informaciones dispersas.

PARA MÁS INFORMACIÓN:

P. Amiet, *L'âge des échanges inter-iraniens: 3500-1700 av.J.-C.* (notas y documentos de Musées de France 11), Éditions de la réunion des musées nationaux, París 1986.
A. Le Brun y F. Vallat, «L'origine de l'écriture à Suse», *Cahiers de la délégation archéologique française en Iran, 8*, 11, 1978.
B. André-Leickman y C. Ziegler (eds.), *Naissance de l'écriture. Cunéiformes et hiéroglyphes*, Réunion des musées nationaux, París, 1982.
M. Powell, «Measure & weights», *Reallexikon der Assyriologie*, 7, 457, 1989.
J. Friberg, «Mathematics», *Reallexikon del Assyriologie*, 7, 531, 1989.
M. Green y H. Nissen (con la col. de P. Damerow y R. Englund), *Zeichenlist der archaischen Texte aus Uruk* (Archaische Texte aus Uruk 2), Gebräuder Mann, Berlín 1987.
D. Schmandt-Besserat, *Before Writing*, 2 vol. , University of Texas Press, Austin, 1992.
J. Ritter, *Mathematiques Mésopotamiennes 4e université d'eté d'histoire des mathématiques*, Irem Lille, 1994.
J. Ritter, «Mésopotamie: une énigme résolue?», *Courrier de l'Unesco, 14*, 1993.
W. Dunham, *El universo de las matemáticas*, Ediciones Pirámide, Madrid, 1996.
R. Dedekind, *¿Qué son los números*, Alianza Editorial, Madrid, 1998.



El teorema de Fermat

Catherine Goldstein

En 1995, Andrew Wiles consiguió que cediera una importante conjetura de las matemáticas actuales, la conjetura de Shimura-Taniyama-Weil. El resultado tuvo una repercusión que rebasó el círculo de profesionales: el establecimiento de la conjetura implica la demostración de un enunciado aritmético de la conjetura de Fermat.

En la primera mitad del siglo XVII, Pierre de Fermat, consejero del Parlamento de Toulouse famoso por sus investigaciones en matemáticas, escribía en el margen de un libro: «*No es posible descomponer un cubo en suma de dos cubos ni una cuarta potencia en suma de dos cuartas potencias ni, en general, ninguna potencia de exponente mayor que 2 en dos potencias del mismo exponente*». En otras palabras, Fermat afirmaba que para n mayor que 2 no hay ningún conjunto de números enteros a, b, c, positivos tales que que satisfagan la ecuación $a^n + b^n = c^n$. Al no haberse encontrado, durante tres siglos y medio, ni prueba general ni contraejemplos, este enunciado de aspecto benigno se convirtió en leyenda. En junio de 1993, Andrew Wiles, investigador británico que trabaja en la Universidad de Princeton (Estados Unidos), anunciaba la demostración de una conjetura central de las matemáticas contemporáneas, la llamada conjetura de Shimura-Taniyama-Weil. La noticia saltó a los periódicos porque desde 1986 se sabe que dicha conjetura implica la aserción de Fermat. El espectacular trabajo de Wiles recurre a todo un arsenal de métodos desarrollados en las últimas décadas y a los resultados de decenas de investigadores. Su resultado, publicado en 1995, no sólo fue el toque final en la resolución de un problema famoso, también sugirió una estrategia par estudiar las relaciones entre distintas ramas de las matemáticas contemporáneas, geometría aritmética, teoría de grupos y de sus representaciones, funciones especiales, y otras muchas.

En el margen de las Aritméticas de Diofanto Pierre Fermat había escrito unas líneas que iban a convertirse en enigma legendario. Su afirmación, traducida a términos modernos, decía que no existían enteros no nulos a, b y c tales que $a^n + b^n = c^n$ y ello cualquiera que fuera el entero n mayor que 2. Además, Fermat indicó que había encontrado una demostración, pero que era demasiado extensa para caber en el margen. (Foto J.-L. Charmet)

Exponentes vencidos. El relato canónico del «teorema» de Fermat antes de junio de 1993 es bien conocido: se van desgranando uno tras otro los exponentes n vencidos. Hay que observar que tan pronto como un exponente ha sido descartado, también quedan descartados todos sus múltiplos. Supongamos, por ejemplo, que el teorema haya sido demostrado para el exponente 13; entonces, la ecuación $(a^m)^{13} + (b^m)^{13} = (c^m)^{13}$ no puede tener por soluciones enteros positivos a, b, y c (pues de otro modo, a^m, b^m y c^m serían enteros positivos solución de la ecuación para el exponente 13). En otras palabras, el teorema de Fermat sigue siendo válido para todos los múltiplos $n = 13\,m$. Basta pues hacer la demostración para el exponente 4 y para todos los exponentes p primos (sin otros divisores que 1 y p) ya que todo entero mayor que 2 es un múltiplo de estos números. Debemos al propio Fermat una prueba para el exponente 4 y a Leonhard Euler, en el siglo XVIII, una prueba para el exponente 3. Los esfuerzos independientes pero complementarios del francés Adrien-Marie Legendre y del alemán J.Peter-Gustav Lejeune-Dirichlet zanjaron el caso de 5 hacia 1825; el de 7 fue resuelto un poco más tarde por el francés Gabriel Lamé. Luego, hacia 1850,

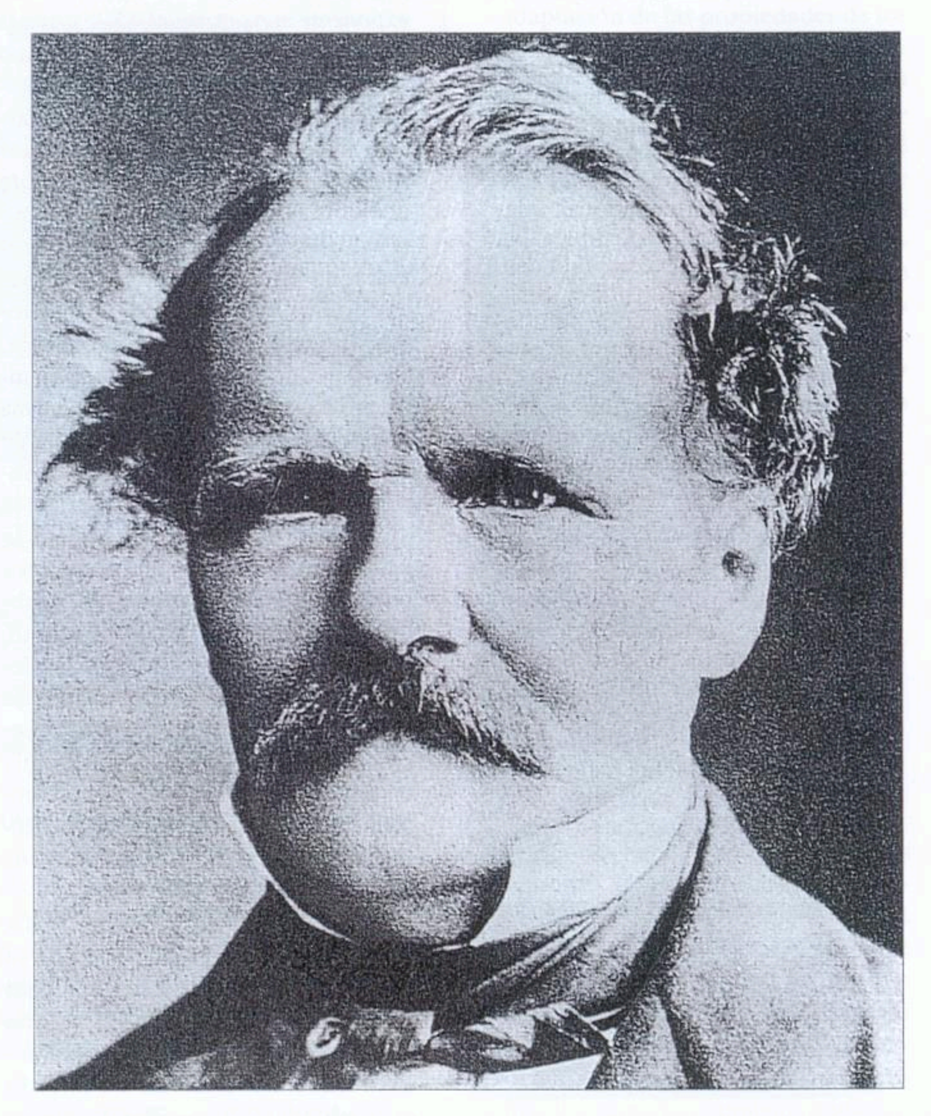
Ernst Eduard Kummer, entonces profesor de la Universidad de Breslau, eliminó de un solo golpe todos los exponentes primos menores que 100 salvo el 37, el 59 y el 67. Y así sucesivamente. En los últimos años se sabía de varios cientos de miles de exponentes primos para los cuales la ecuación $a^p + b^p = c^p$ (y por tanto todas las ecuaciones con exponentes n múltiplos de p) carece de solución en términos de números enteros positivos a, b y c.

Una demostración fantasma. Este relato tradicional se adereza a veces con otros resultados pertinentes, parciales con respecto a las soluciones eliminadas, pero que tienen el interés de abarcar familias potencialmente infinitas de exponentes. Por ejemplo, el de Sophie Germain, a principios del siglo XIX, se refería a las soluciones a, b, c, no divisibles por el exponente p. Pero la primera prueba que excluía estas soluciones para una infinidad de exponentes primos no fue dada hasta 1985 por E. Fouvry, de la Universidad de Burdeos. Por su parte, el alemán G. Faltings (medalla Fields 1986) demostró en 1983 que para todo exponente mayor igual a 5 la ecuación de Fermat tenía como máximo un número finito de soluciones sin divisores comunes. A estos avances significativos cabría añadir a voluntad un conjunto heterogéneo de simplificaciones, teoremas conexos y tentativas fallidas.

Con la desaparición de la solución de Fermat nacieron las esperanzas tenaces en una demostración elemental de la conjetura

Una tal sucesión de nombres y fechas se presta a confusión. Convierte el teorema de Fermat en el tema central de los esfuerzos de todos los matemáticos citados, como lo fue de cientos de aficionados que hoy siguen enviando «pruebas» (falsas) del teorema a las Academias de ciencias o a las universidades. Por supuesto, la fama creciente del problema fomentó su tratamiento con otros puntos de vista, otras técnicas y otros procedimientos de cálculo; en muchos casos, constituyó una motivación personal. Pero contrariamente a lo que a menudo se lee, la observación de Fermat no ha sido tanto un estímulo fundamental de desarrollo de la teoría de números como la compañera y el testigo fiel de sus cambios y sus múltiples componentes. Así, por no citar sino los trabajos basados en los avances más recientes, Fouvry utilizó refinadas estimaciones analíticas, el resultado de Faltings es una consecuencia de difíciles problemas de geometría algebraica y la demostración propuesta por Wiles navega entre la geometría aritmética de curvas y la teoría de representaciones. Los grandes «progresos», en términos de exponentes, que ritmarían por ejemplo los trabajos de Fermat, Kummer y Wiles, son la marca de programas individuales y colectivos de mayor envergadura, cada uno de los cuales daría motivo para una o varias historias propias. Enfocaremos nuestro proyecto en estos tres momentos, sumergiéndonos cada vez en un mundo matemático específico para tratar de captar sus motivaciones, sus métodos y su especial lenguaje.

Empecemos por la leyenda. A su famoso enunciado, Fermat le añadió un comentario más famoso todavía: «He conseguido una maravillosa demostración. La estrechez del margen no puede contenerla». Pero nunca se ha encontrado ninguna prueba general, verdadera o falsa, en los papeles de Fermat, ni después de su muerte, ni



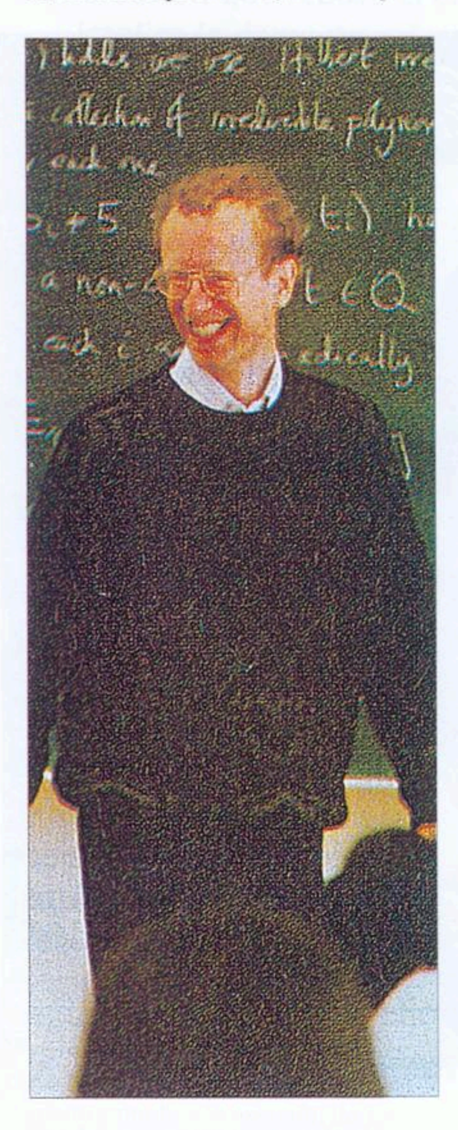
La importancia creciente de la teoría de números en matemáticas en el siglo XIX contribuyó a dar fama al teorema de Fermat, pero también a convertirlo en accesorio. Hacia 1850, el alemán E.E. Kummer consiguió demostrarlo para exponentes inferiores a 100, salvo tres, al término de un trabajo sobre las «leyes de reciprocidad». Desde entonces se han obtenido otros resultados importantes y recientemente se han elaborado varios enfoques prometedores. En 1986, por ejemplo, K. Ribet, de la Universidad de California en Berkeley, demostró que el teorema de Fermat deriva de una importante conjetura de las matemáticas contemporáneas, la conjetura de Shimura-Taniyama-Weil. (Foto Bildarchiv Preussischer Kulturbesitz)

más tarde, cuando los investigadores emprendieron la edición de sus obras a finales del siglo XIX. De esta novelesca desaparición nació la esperanza de poder dar una demostración elemental que recurriera únicamente a los instrumentos matemáticos de que disponía Fermat.

La «pereza» de Fermat. La disipación de este misterio no incumbe tanto a las matemáticas como a la historia y más especialmente al estudio de los escritos científicos de Fermat. La observación en cuestión es una nota personal, probablemente escrita antes de 1840, cuyo autógrafo no ha sido encontrado. En su correspondencia, Fermat plantea numerosos problemas aritméticos, generales o no, o anuncia su demostración. Menciona en varias ocasiones que un cubo no puede ser la suma de dos cubos y el enunciado análogo para la cuarta potencia. En cambio nunca alude al caso general (para todo n mayor que 2) del teorema. Cuando en 1659 Fermat envió a su amigo Pierre de Carcavi un balance global de sus investigaciones sobre los números, todavía citaba sólo el caso de los cubos (así como un resultado próximo al de las cuartas potencias). Salvo hallazgo de nuevos documentos, pues, se puede concluir, con el historiador Jean Itard, que *«Fermat nunca estuvo en posesión de una prueba de su teorema mayor para un exponente mayor o igual a 5»*[1] y que la nota al margen traducía un entusiasmo de neófito que pronto se atemperó.

Las pruebas para los cubos y las cuartas potencias bastan para despertar nuestro interés, ya que carecemos de redacción completa de ellos, como tampoco la tenemos de los demás resultados de Fermat. Acerca de las cuestiones que debieron de apasionar a los matemáticos de su época, René Descartes el primero, por ejemplo la búsqueda de tangentes a curvas o el cálculo de áreas, Fermat debió de redactar soluciones, responder a objeciones y precisar la extensión de la aplicación de sus trabajos. Pero la aritmética, aunque tenía algunos adeptos apasionados, era también objeto de desdén. Era fastidiosa e inútil, sobre todo cuando se trataba de demostrar que un problema carecía de solución... Fermat era «el más perezoso de los hombres», según propia confesión, y abrumado por sus ocupaciones profesionales, no mostró demasiado interés en explicitar sus pruebas en este campo.

Su actividad y su proyecto, por tanto, han de reconstruirse partiendo de su correspondencia y de notas análogas a la que nos interesa. Estas últimas acompañan las Aritméticas de Diofanto (hacia el siglo III de nuestra era), editadas y traducidas del griego al latín y comentadas en 1621 por un erudito francés, Claude-Gaspard Bachet de Méziriac. El conjunto, incluidas las notas de Fermat, lo volvió a publicar en 1670, después de su muerte, su hijo Samuel. Se trata de una colección de problemas en los que se buscan números, en general enteros o fraccionarios, sometidos a determinadas condiciones: por ejemplo, tres números tales que las sumas de dos de ellos y de su producto sean cuadrados. O también dos números cuadrados cuya suma sea un cuadrado.



En junio de 1993, A. Wiles esbozó la demostración de una parte importante de la conjetura, una parte que bastaría para establecer el teorema de Fermat. Estos trabajos prometedores fueron rápidamente confirmados y valieron al matemático numerosos premios.
(Foto P. Goddard/SPL/Cosmos)

Eficacia algebraica. Todas la civilizaciones han conocido soluciones a este último problema, como 25 = 16 + 9 ($5^2 = 4^2 + 3^2$) o también 169 = 144 + 25 ($13^2 = 12^2 + 5^2$). Hay una infinidad de tales números: gracias al llamado «teorema de Pitágoras», ello equivale a la existencia de una infinidad de triángulos rectángulos de lados enteros. Fue a raíz de esta cuestión sobre las sumas de dos cuadrados cuando Fermat advirtió que en cambio no había ningún cubo que fuera suma de dos cubos o ninguna cuarta potencia que fuera suma de dos cuartas potencias. Por lo demás, la idea de esta generalización no era nueva. El caso de los cubos, por lo menos, era famoso ya entre los matemáticos de lengua árabe. El médico y gran científico Avicena lo menciona incluso en el siglo XI en un tratado de filosofía.[2]

Para su autor, el teorema, lejos de ser un fin en sí mismo, no era sino un ejemplo que permitía ilustrar la potencia del método

Para Fermat, este tipo de enunciados no tenían un interés aislado; formaban parte de un programa y estaban destinados a ilustrar la potencia de un método. *Las Aritméticas* de Diofanto, en efecto, aparecen en numerosas bibliotecas científicas del Renacimiento como una reserva de tests destinada a probar la eficacia del álgebra, elaborada en los países islámicos, que los matemáticos occidentales empezaban entonces a asimilar y desarrollar. Las condiciones sobre números desconocidos que allí se encuentran se traducen fácilmente en relaciones algebraicas. Competían varias maneras de tratar las ecuaciones e incluso varios simbolismos; cada autor alababa las ventajas teóricas o prácticas de los suyos.

Pero las manipulaciones algebraicas dependen poco, y en esto consiste su generalidad, del carácter particular

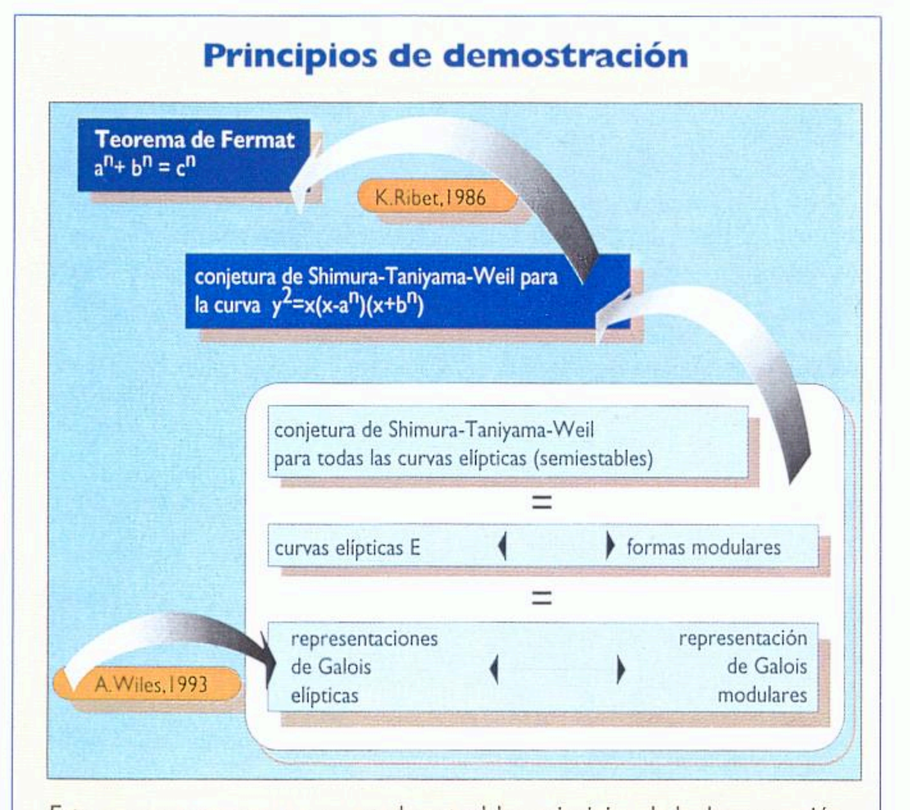
de los números sobre los que se ejercen. Igual valen los números enteros que los irracionales y las notaciones no distinguen unos de otros. Fermat quería sacar partido de la potencia de estas técnicas y a la vez garantizar la especificidad de la teoría de los números, que para Fermat era la teoría de los enteros. Utilizó a tal fin un método llamado de descenso *infinito*, que consistía en crear, a partir de una solución entera de un problema, otra solución formada por enteros estrictamente más pequeños, por medio de transformaciones algebraicas.

Fue este método el que el matemático dijo haber utilizado para los exponentes 3 y 4 (sólo disponemos de pistas para el caso de 4). Lejos de ser un fin en sí mismo, el que sería el teorema más famoso de Fermat no parecía para su autor otra cosa que un ejemplo de su programa, que no fue proseguido tal cual. Pero los enunciados de Fermat (con la excepción que sabemos) serían demostrados antes del siglo XIX e integrados a otras preocupaciones. Así ocurrió, por ejemplo, con la representación de los números como sumas de cuadrados y de potencias diversas.[3]

En aquella época, dos siglos después de la lectura por Fermat de las Aritméticas, la teoría de números se transformó esencialmente, a la vez en su temática y en su estatuto. Dejó de ser marginal y empezó a ocupar un lugar privilegiado en las universidades alemanas.

Enteros ciclotómicos. Fue parcialmente responsable de este viraje una autonomía creciente de los profesores, que contrastaba por ejemplo con el caso de la Escuela politécnica de Francia, donde la presión de los ingenieros orientaba las investigaciones hacia disciplinas más directamente aplicadas.[4] Apareció una nueva línea de investigación: la extensión o la adaptación de las propiedades de los enteros usuales —tales como su factorización en producto de números primos o su representación como suma de determinadas potencias— a otras clases de números. Algunos fenómenos relativos a los enteros usuales podían comprenderse mejor ampliando el campo de trabajo. En este marco hay que ubicar los importantes artículos que Kummer dedicó al teorema de Fermat entre los años 1847 y 1857, que le valieron un premio de la Academia de ciencias de París.

Principios de demostración



Este esquema resume, en su estado actual, los principios de la demostración del teorema de Fermat. En 1986, K. Ribet demostró que el teorema de Fermat deriva de la conjetura de Shimura-Taniyama-Weil (STW) para la curva elíptica de ecuación $y^2 = x(x - a^n)(x + b^n)$. Los números enteros a, b y n ($n > 2$) son los que intervienen en la relación $a^n + b^n = c^n$, objeto del teorema de Fermat. Ribet demostró que de existir dichos números enteros la correspondiente curva elíptica violaría la conjetura STW; por tanto, si ésta es verdadera, dichos números no existen y el enunciado de Fermat es verdadero. La conjetura STW predice la existencia de una correspondencia precisa entre el conjunto de las curvas elípticas y el conjunto de las funciones o «formas» llamadas modulares. En 1993 Wiles dio un esquema de demostración de la conjetura STW para el caso de las llamadas curvas elípticas semiestables, caso particular que basta para establecer el principio de Fermat. Su método consiste en establecer una correspondencia entre unos objetos matemáticos asociados a las curvas elípticas (las «representaciones de Galois elípticas») y unos objetos asociados a las formas modulares (las «representaciones de Galois modulares»). Para ello, Wiles se basa en los trabajos de numerosísimos matemáticos contemporáneos.

El gran hallazgo de Kummer consistió en introducir unos nuevos «números ideales» en los cuales vuelven a descomponerse los enteros ciclotómicos

Kummer, que llegaría a ser una de las grandes autoridades de la Universidad de Berlín, se interesaba entonces por los números que actualmente se denominan *enteros ciclotómicos*. Se trata de números enteros formados por sumas de enteros usuales y de raíces enésimas de 1, siendo estas últimas los números complejos que satisfacen ζ verifica $\zeta^n = 1$. Pertenecen a esta categoría los «enteros de Gauss», números de la forma $a + ib$, donde a y b son enteros e $i = \sqrt{-1}$. Los enteros de Gauss son ciclotómicos, pues i es una raíz cuarta de la unidad: ($i^4 = 1$). Uno de los objetivos de Kummer, continuador de las investigaciones de C.F. Gauss o C.G.J. Jacobi, consistía

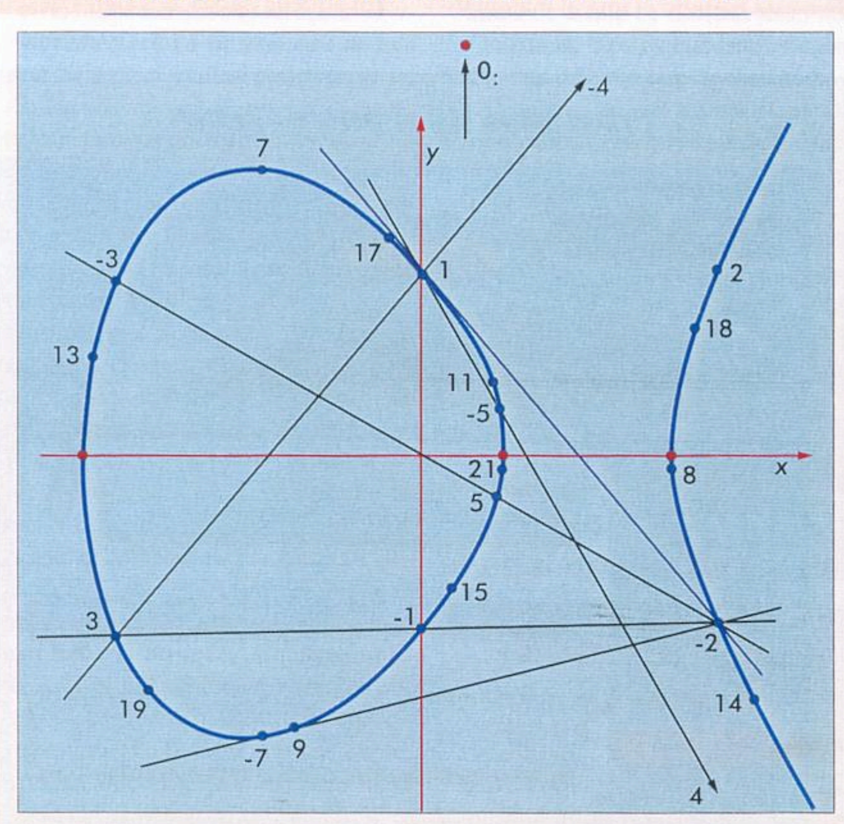
en poner de manifiesto *leyes de reciprocidad* entre números primos. La más simple de estas leyes afirma que si p y q son dos números primos distintos, de los cuales al menos uno es igual a $4m + 1$, (como 5, 13, 17, etc.), entonces p es un cuadrado salvo un múltiplo de q exactamente cuando q es un cuadrado salvo un múltiplo de p. Este tipo de leyes tienen interés práctico —el cálculo de cuadrados salvo múltiplos enteros interviene, por ejemplo, en problemas de acústica—, y teórico, ya que establecen las relaciones entre los números primos, elementos de construcción de los números enteros. Se sabía ya que los enteros de Gauss proporcionaban un marco adecuado para discutir algunas de estas leyes, por lo que Kummer emprendió un estudio de los enteros ciclotómicos generales para acceder a potencias distintas de los cuadrados. Sus trabajos y los archivos con ellos relacionados han sido prolijamente analizados por el historiador estadounidense Harold Edwards.[5]

El trabajo de Wiles se basa en resultados publicados en más de sesenta artículos y libros que datan de hace menos de treinta años

Los enteros ciclotómicos tienen muchas propiedades en común con los enteros usuales. Se suman y multiplican entre sí e incluso se descomponen en producto de enteros ciclotómicos indescomponibles, que desempeñan un papel análogo a los números primos de la aritmética usual. Después de realizar cálculos extensivos, Kummer constató sin embargo que esta descomposición no era en general única, contrariamente al caso ordinario. Su gran hallazgo fue la introducción de unos nuevos números, llamados «números ideales», en los cuales vuelven a descomponerse los números ciclotómicos, y la recuperación, por esta vía, de la unicidad de la descomposición. La posibilidad de aplicar estas nuevas ideas al problema de Fermat procede de las descomposición $a^p + b^p = (a + b)(a + \zeta b)(a + \zeta^2 b)\ldots(a + \zeta^{p-1} b)$, donde $\zeta \neq 1$ es una raíz compleja p-ésima de la unidad, es decir, un número que verifica $\zeta^p = 1$ (por ejemplo, $\zeta = \cos(2\pi/p) + i\ \mathrm{sen}(2\pi/p)$). Para quien esté interesado en las distintas tentativas de demostrar el teorema de Fermat hay que decir que Kummer no fue el primero en utilizar los números complejos a título auxiliar. Lamé y Augustin Cauchy llegaron incluso a abordar el problema, aproximadamente en la misma época, por medio de estas mismas raíces p-ésimas de la unidad. Pero Kummer, basándose en el conocimiento general de los enteros ciclotómicos y en la unicidad de la factorización proporcionada por los números ideales, fue capaz de demostrar el teorema de Fermat para todos los exponentes p que verificaban determinadas condiciones técnicas ligadas a las estructuras de los enteros ciclotómicos y de sus números ideales. En particular, Kummer verificó la condición para todos los números primos inferiores a 100, excepto tres (37, 59 y 67).

Estas condiciones sobre p fueron modificadas más tarde, al precio de importantes trabajos teóricos que tam-

Suma de los puntos de una curva elíptica



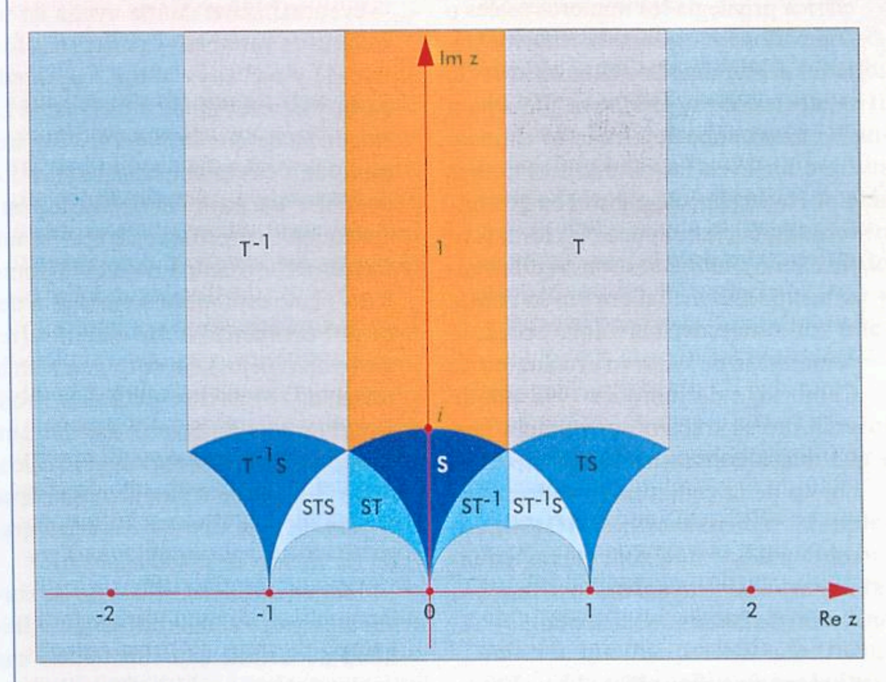
La demostración del teorema de Fermat pasa por la demostración de la conjetura de Shimura-Taniyama-Weil, que a su vez hace intervenir unas curvas llamadas elípticas. Es posible definir una «suma» entre los puntos de dicha curva. La figura muestra el ejemplo de la curva elíptica E de ecuación $y^2 = x^3 - x + 1/4$. Pertenece a E el punto de coordenadas racionales $P = (0, 1/2)$, llamado aquí simplemente 1 (por 1.P). La tangente en dicho punto vuelve a cortar la curva en el punto llamado -2. Su simétrico con respecto al eje Ox es, por definición, el resultado de sumar P + P y se representa aquí por 2 (por 2P). Más generalmente, la suma $P_1 + P_2$ de dos puntos distintos se define así: se prolonga la secante que une P1con P2 y se toma el simétrico del tercer punto de intersección de la curva con esta secante. El cero de esta suma es el punto del infinito llamado 0_∞. Se han indicado en el esquema los primeros múltiplos (2P, 3P, 4P, etc.) del punto P anteriormente definido. Se puede demostrar que todos los puntos de coordenadas racionales de la curva son múltiplos de P. Hay tres puntos Q no nulos que verifican 2Q = 0, aquellos para los cuales y es nulo y x es solución de $x^3 - x + 1/4 = 0$. Se trata de los tres puntos de intersección de E con el eje Ox (en rojo). Con el punto 0_∞ forman el conjunto E[2] mencionado en el artículo. (Tomado de R. Hartshorne, *Algebraic geometry*, Springer, 1977)

bién facilitaron su verificación numérica; los ordenadores extendieron el campo de los exponentes accesibles. Hasta los últimos años, este enfoque siguió siendo fundamental para explorar concretamente nuevos exponentes. Es curioso que Kummer considerara el teorema de Fermat como una «curiosidad» (*sic*), mientras que las leyes de reciprocidad le parecían el pináculo de la teoría de números. La exploración de estas últimas, y más generalmente la transferencia de las propiedades aritméticas a distintos campos mucho más generales que los enteros ciclotómicos, constituyen todavía un campo de investigación muy activo cuyas ramificaciones se han multiplicado (teoría algebraica de números y cuerpos de funciones, teoría K, estructuras algebraicas, lógica, etc.)

Movilización. Pero aunque el paisaje aritmético de finales del siglo XX ha heredado algunos rasgos institucionales y matemáticos de la teoría de números alemana del siglo anterior, otros son mucho más recientes. Así, Wiles cita en su trabajo más de sesenta artículos y libros de hace menos de treinta años, una parte importante de los cuales, que desempeña un papel fundamental en sus argumentos, tiene menos de diez años. La impresionante lista de personas cuyos resultados ha utilizado directamente o le han servido de inspiración abarca decenas de especialistas de todo el mundo. Por supuesto, semejante movilización técnica no tenía por único objetivo probar el teorema de Fermat, aunque las conexiones con este último fueran al final un acicate individual. Como hemos dicho al principio, lo que Wiles explicó en junio de 1993 fue una estrategia para probar un caso importante de la llamada conjetura de Shimura-Taniyama-Weil que llamaremos de ahora en adelante STW para abreviar (véase el recuadro «Principios de demostración»). Esta conjetura se remonta a los años 1960 y predice entre ciertas curvas definidas por una ecuación de tercer grado, las «curvas elípticas» y unas funciones específicas, llamadas «modulares», una relación análoga a la del círculo con las funciones circulares, coseno o seno. Volveremos a la conjetura y al trabajo de Wiles pero explicaremos en primer lugar cómo implica el teorema de Fermat.

De una conjetura a la otra. Supongamos que éste sea falso, es decir, que existan enteros positivos a, b y c no nulos que cumplan $a^p + b^p = c^p$ para un número primo p mayor o igual a 2. Consideremos entonces la curva plana E_p de ecuación (en x e y, elegidas como coordenadas del plano) $y^2 = x(x - a^p)(x + b^p) = x^3 + (b^p - a^p)x^2 - (ab)^p x$. Esta curva es una curva elíptica cuyos coeficientes están definidos a partir de la ecuación de Fermat. Se demuestra entonces que su existencia es incompatible con la conjetura de Shimura-Taniyama-Weil. En otros términos, el establecimiento de esta última implica que la curva asociada a dichos coeficientes no puede existir y que por tanto el teorema de Fermat es verdadero. Un argumento similar permitiría tratar gran cantidad de situaciones análogas a la de Fermat. Por ejemplo, se demuestra también que la ecuación $a^p + 17b^p = c^p$ carece de soluciones enteras no nulas para $p \neq 17$, así como otros resultados del mismo tipo. Esta relación entre la conjetura STW y la ecuación de Fermat no se debe a Wiles; las curvas E_p fueron estudiadas en los años 1970 por Y. Hellegouarch, entonces en la Universi-

Funciones Modulares



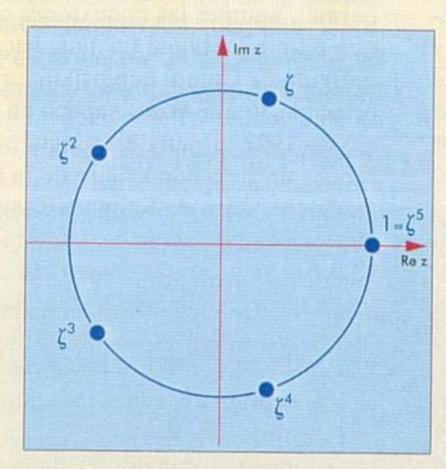
La conjetura de Shimura-Taniyama-Weil establece una correspondencia entre curvas elípticas y «funciones modulares». Las funciones modulares g de «nivel» N están caracterizadas por ciertas propiedades de invariancia. Debe verificarse g(z') = g(z) para todo número complejo z, donde z' = (az + b)/(gz + d), y ello para enteros a, b, g, d tales que ad - bg = 1 y tales que g sea un múltiplo de N. La figura muestra el plano complejo de z; la zona amarilla representa el «dominio fundamental» de las funciones modulares de nivel N = 1. Cada región delimitada se deduce de este dominio llamado «1» por una transformación especial de tipo z→(az + b)/(gz + d), con ad - bg = 1. Por ejemplo el dominio T se obtiene por traslación de 1 (es decir, z→ z + 1), el dominio S, por z→ -1/z, TS por z→ 1-1/z, etc. Agotando todas las transformaciones se recubriría todo el semiplano superior. Dado que las funciones modulares de nivel N = 1 son invariantes por estas transformaciones, están completamente determinadas por sus valores en la región 1. Para los demás valores del nivel N hay menos transformaciones permitidas (pues N ha de ser un divisor de g) y la función queda determinada por sus valores en una reunión finita de dichas regiones llamada «dominio fundamental» del conjunto de transformaciones. Para N = 2, este dominio comprende tres regiones (por ejemplo 1, S y ST^{-1}); para N = 4, seis regiones; para N = p primo, p + 1 regiones.

dad de Besançon.[6] G. Frey, de la Universidad de Sarrebruck, las devolvió a un primer plano al sugerir que estas curvas no verificaban la conjetura STW.[7] La prueba completa, en modo alguno evidente, está ligada a otras conjeturas y a los trabajos de J.P. Serre, del Collège de France.[8] La demostración del vínculo entre la conjetura STW y la conjetura de Fermat fue dada en 1986 por K. Ribet, en la Universidad de California, en Berkeley, lo que le valió el premio Fermat de 1989, concedido por la Universidad Paul-Sabatier de Toulouse.[9]

> Las recientes aproximaciones al teorema estaban marcadas por una fuerte interacción entre teoría de números y geometría algebraica

Representación de Galois

La ecuación $z^5 = 1$ tiene cinco soluciones, entre las cuales figura la solución racional $z = 1$. las otras cuatro raíces son el número complejo $\zeta = \exp(2i\pi/5) = \cos(2\pi/5) + i\,\text{sen}(2\pi/5)$ y sus potencias $\zeta^2, \zeta^3, \zeta^4$, que corresponden a los ángulos $4\pi/5$, $6\pi/5$ y $8\pi/5$, respectivamente. Existen *a priori* 120 permutaciones posibles entre estas cinco raíces, pero muchas menos que respeten sus relaciones algebraicas. Para tales permutaciones, $z = 1$ es necesariamente fijo y las otras cuatro raíces deben ser permutadas entre sí; además, el efecto de la permutación sobre ζ determina el de las otras raíces por elevación a la correspondiente potencia. Quedan sólo, por tanto, cuatro posibilidades, según que ζ quede fijo o sea transformado en ζ^2, ζ^3 o ζ^4. Por ejemplo, si ζ se transforma en ζ^3 se obtiene la permutación $(1' \to 1, \zeta \to \zeta^3, \zeta^2 \to \zeta^6 = \zeta\ \zeta^3 \to \zeta^4, \zeta^4 \to \zeta^2)$. La «representación de Galois» indicaría, para cada transformación algebraica, la permutación correspondiente a su acción sobre las raíces. La conjugación compleja, por ejemplo, cambia ζ en ζ^4. Este efecto puede codificarse simplemente mediante un número comprendido entre 1 y 4 que es la potencia en la que queda transformado ζ; la representación de Galois es la familia constituida por estos códigos.



El caso de las curvas elípticas y de las representaciones de Galois asociadas es idéntica, con los puntos de E[n] sustituyendo las raíces de la unidad. La diferencia estriba en que hace falta entonces un cuadro de cuatro números para codificar cada transformación.

No es paradójico sustituir una conjetura aparentemente imposible de probar (en este caso el «teorema» de Fermat) por otra, la de Shimura-Taniyama-Weil, más «fuerte» y por tanto en principio más difícil de demostrar. En primer lugar, los especialistas tenían mayor confianza en la conjetura STW, conectada a otros muchos resultados de las matemáticas, que en un teorema aislado como el de Fermat, por famoso que fuera.

Además, el carácter muy técnico de la conjetura STW constituye una baza, ya que suministra de entrada unas herramientas con que trabajar. Las recientes aproximaciones al teorema de Fermat tienen en mayor o menor grado este carácter técnico.[10] También están marcadas por la fuerte interacción entre teoría de números y geometría algebraica que caracteriza algunos temas aritméticos, sobre todo desde los años 1960. La «geometría» en cuestión fue desarrollada en los trabajos de A. Grothendieck (medalla Fields 1966) y de sus sucesores. Al precio del despliegue de nuevos instrumentos, dicha geometría es capaz de captar los dominios donde se definen sus objetos familiares, curvas, superficies y las generalizaciones de éstas, y tratarlas uniformemente. Existe así una geometría «aritmética», que permite trabajar con los enteros, mientras que la geometría clásica privilegia los números reales o complejos. Lo veremos con mayor claridad al examinar la conjetura STW y la demostración sugerida por Wiles para una amplia familia de curvas elípticas (que incluyen las paradójicas curvas E_p). Hay tres protagonistas en esta historia: las curvas elípticas, las formas (*o funciones) modulares y las representaciones de Galois*. Tal era por lo demás el título, sugestivo pero equívoco, de la conferencia de Wiles en el coloquio de Cambridge de junio de 1993 donde presentó su trabajo por primera vez. La relación entre las dos primeras es objeto de la conjetura STW, mientras que las representaciones de Galois, crisol donde se comparan curvas y funciones, constituyen un instrumento crucial de las demostraciones.

Suma de puntos. Empecemos pues por las curvas elípticas, al menos por aquellas que aquí nos conciernen. Se trata de curvas planas E lisas (es decir, sin nudos ni picos) definidas por una ecuación en x e y de tercer grado con coeficientes racionales que podemos —eventualmente con la ayuda de un cambio de variables— poner en la forma: (1) $y^2 = x^3 + a_2 x^2 + a_4 x + a_6$ donde a_2, a_4 y a_6 son números racionales. La exigencia de que la curva sea lisa impone una cierta condición sobre los coeficientes; para las curvas E_p esta condición es precisamente que los números a, b, c sean no nulos. Un punto P del plano está sobre la curva E si tiene por coordenadas dos números (reales o complejos) x e y que cumplen la relación (1). Las curvas elípticas, prolijamente estudiadas, intervienen en contextos tan variados como el cálculo de la longitud de un arco de elipse —y de ahí su nombre— o la criptografía (la ciencia de la codificación).

El rasgo más característico de las curvas elípticas, y el más útil, es la posibilidad de definir una suma entre sus puntos (véase recuadro «Suma de puntos de una curva elíptica»). Este notable fenómeno tiene su origen en la llamada «construcción de secantes y tangentes»: si P_1 y P_2 son dos puntos de una curva, la recta secante que pasa por ambos corta en general a la curva

en otro punto P_3. Asimismo, la tangente a la curva en un punto P la vuelve a cortar en un nuevo punto P'.
La curva es simétrica con respecto al eje de abscisas y se define simplemente el opuesto de un punto P (o –P) como el simétrico de P con respecto a dicho eje. La suma de dos puntos distintos P_1 y P_2, $P_1 + P_2$ se define como el opuesto (esto es, el simétrico con respecto al eje Ox) del punto P_3; por consiguiente, tres puntos de la curva tienen una suma nula cuando están alineados. Se define también la suma de dos puntos coincidentes P + P, llamada 2P, como el opuesto de P'. Es fácil calcular las coordenadas de la suma de dos puntos (iguales o distintos) a partir de las coordenadas de los puntos. Para evitar excepciones (en particular, la tangente en un punto, si es vertical, no corta la curva) se añade a los puntos de la curva un punto en el infinito, el punto de fuga de las verticales. Dicho punto hace las veces de «cero» para la suma y lo llamaremos 0 u 0_∞.

Aunque hace intervenir objetos geométricos o analíticos, la conjetura STW expresa un comportamiento aritmético

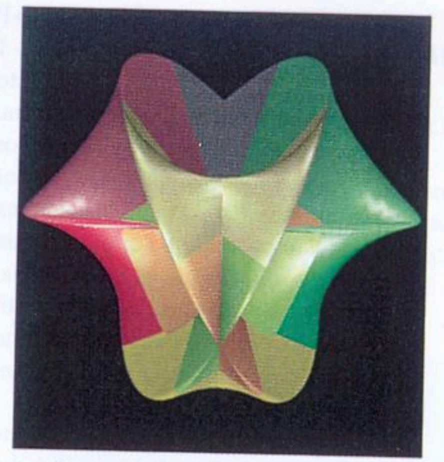
Esta «suma» de puntos tiene propiedades inesperadas: se pueden sumar los puntos en cualquier orden y agruparlos arbitrariamente para efectuar la suma paso a paso. En otras palabras, los puntos de E tienen una estructura de grupo conmutativo.
Contrariamente a la suma usual de enteros, sin embargo, un múltiplo (no nulo) de un punto puede ser nulo. Así, los puntos de la curva situados sobre el eje Ox son sus propios simétricos; verifican pues P = –P, es decir, 2 P = 0. Hay tres puntos de esta clase; sus abscisas x son las tres soluciones de $0 = x^3 + a_2x^2 + a_4x + a_6$. El conjunto de estos tres puntos y del punto O∞ se llama tradicionalmente E[2] ; es estable por la suma: la suma de dos de tales puntos, o el opuesto de uno de ellos pertenecen también a E[2] (fig. 4).

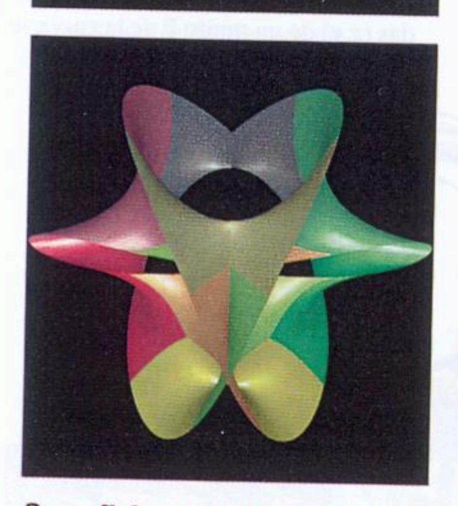
Parametrización de una curva. Más generalmente, cabe definir para la curva unas subestructuras finitas E[n] formadas por puntos tales que nP = P +P +...+ P($n$veces) = 0; E[n] contiene n^2 - 1 elementos, a los que se añade el punto O_∞ y también es estable bajo la suma. Obviamente, la mayoría de los puntos de E no pertenecen a ninguno de estos subconjuntos E[n]. Pero éstos juegan un papel importante, ya que el estudio de algunas de sus simetrías bastan para reconstruir casi toda la curva elíptica. Como veremos éste es el camino que toma Wiles para abordar la demostración.
Pasemos ahora al segundo ingrediente fundamental, las *funciones modulares*. Como ya hemos dicho, estas funciones tienen rasgos en común con las funciones seno y coseno. La función cos($2\pi t$) está definida para todo número complejo t, es regular (por ejemplo, desarrollable en serie de potencias de t) y periódica de período 1. Las funciones modulares g admiten también desarrollos en series, pero sólo están definidas en la mitad superior del plano complejo, es decir, para $t = t1 + it_2$, con $t_2>0$ (véase recuadro «Funciones modulares»); en cambio, sus propiedades de periodicidad son más fuertes. En el caso más simple, deben verificar, para un entero N > 1 fijado llamado «nivel» de la función: $g((\alpha t + \beta)/(\gamma t + \delta)) = g(t)$ para todos los enteros α, β, γ, δ —con γ divisible por N— tales que αβ -γβ = 1.
Para α = β = δ = 1 y γ = 0 se encuentra por ejemplo la periodicidad 1: $g(t + 1) = g(t)$.
A finales de los años 1950, el matemático japonés Yutaka Taniyama había vislumbrado una relación entre tales funciones y las curvas elípticas, una relación que fue precisada, estructurada y difundida en la comunidad matemática de los años 1960 gracias a varios trabajos, en particular los del japonés Goro Shimura y el francés André Weil. Hay varias formulaciones de la conjetura de Shimura-Taniyama-Weil. Cada una pone énfasis en un aspecto particular, con ventajas propias, y Wiles las utiliza todas. Mencionemos aquí tres. La primera es la más simple de enunciar, la segunda se presta bien a enunciados explícitos y la tercera, en la que intervienen las representaciones de Galois, es el punto de partida del trabajo de Wiles.
En su forma más simple, la conjetura STW predice que los puntos de una curva elíptica cuya ecuación contenga coeficientes racionales pueden ser parametrizados por funciones modulares. En otras palabras, las coordenadas (x,y) de un punto P de la curva se

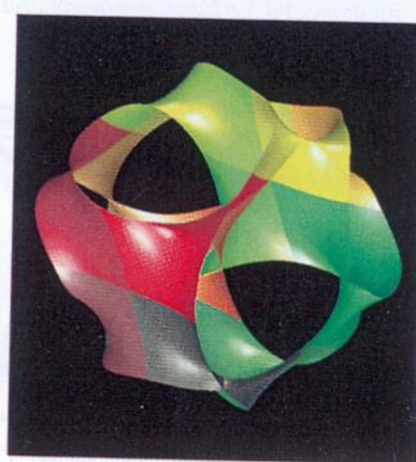


escriben $x = g(t)$, $y = h(t)$, donde g y h son funciones modulares de «nivel» N bien determinado.
Esta relación es más sutil de lo que parece. En efecto, siempre es posible parametrizar los puntos de una curva plana lisa, definida por una ecuación de tercer grado de coeficientes complejos, mediante funciones periódicas. Pero el hecho de que aquí convengan funciones modulares está ligado al hecho de que los coeficientes de la ecuación [1] son racionales. Pese a hacer intervenir objetos geométricos (las curvas elípticas) o analíticos (las funciones modulares), la conjetura STW expresa, pues, un comportamiento aritmético.

La tercera protagonista. Las funciones f y g dependen de las coordenadas elegidas para representar la curva y es interesante disponer de una formulación más cómoda para las verificaciones; esta segunda formulación es una versión diferencial de la anterior. Afirma que cabe asociar a la curva E una serie:







Superficies en 4 D asociadas con las ecuaciones de Fermat N = 3 y N = 4. Center for Innovative Computer Applications: «Project Fermat Images» (www.cica.indiana.edu/projects/Fermat/images.html)

$(t) = e^{2i\pi t} + c_2 e^{4i\pi t} + c_3 e^{6i\pi t} + ...$, definida en el semiplano complejo superior, con coeficientes c_i enteros, tal que la forma diferencial $f(t)dt$ sea invariante bajo las mismas transformaciones, ligadas al entero N, que antes. Estas series se llaman *formas modulares de nivel N* —la terminología exacta, que poco nos importa aquí, es «forma modular parabólica de peso 2 y nivel N»—. Ahora bien, se puede establecer una correspondencia muy precisa entre curvas elípticas y formas modulares: para p primo, el coeficiente c_p de la serie f es igual a p menos el número de pares de enteros (x,y) comprendidos entre 0 y p - 1 tales que el valor de $y^2 - x^3 - a_2x^2 - a_4x - a_6$ sea un múltiplo de p; el nivel N se expresa a partir de la ecuación de la curva. Por ejemplo, para la curva $y^2 = x^3 - x + 1$ y $p = 3$, hay seis pares (x,y) de esta clase que son (0,1), (0,2), (1,1), (1,2), (2, 1), (2,2). El coeficiente c_3 de la forma f asociada es, por tanto, 3 - 6 = -3.

Por otra parte, todas las formas modulares de nivel N como las anteriores se obtienen como combinación lineal de un número finito calculable de ellas. También se puede demostrar que este número es igual a 0 para N < 10 o N = 12 (en otros términos, no hay formas modulares adecuadas para estos valores de N), que vale 1 para N = 11, 14, 19, ..., que vale 2 para N = 37, etc. Ello permite precisar un poco por qué la conjetura STW implica el teorema de Fermat. Las curvas elípticas E_p, de existir y verificar la conjetura STW, deberían tener como nivel N de las formas modulares asociadas el producto de los divisores primos del número *abc*. El punto delicado que demostró Ribet en 1986 es que si nos restringimos al pequeño subconjunto $E_p[p]$ de los puntos Q de E_p tales que $pQ = 0$, el nivel N de las funciones y formas modulares necesarias puede reducirse hasta 2. Pero, como acabamos de decir, no hay forma adecuada para el nivel 2, de donde la contradicción entre la existencia de las curvas E_p y la conjetura STW. Nuestra tercera formulación de la conjetura STW permite introducir las terceras protagonistas de esta historia, las *representaciones de Galois* (véase recuadro, pág. 22). En muchas ramas de las matemáticas y de la física, es habitual estudiar un espacio a partir de sus simetrías estructurales y, recíprocamente, comprender mejor un grupo complicado interpretándolo como un grupo de simetrías de un espacio conocido; se habla entonces más bien de representaciones del grupo. Consideremos, por ejemplo, el grupo S_4 de permutaciones entre cuatro letras A, B, C, D. Cabe interpretado como el grupo de determinadas simetrías del cubo. Basta para ello designar cada una de las diagonales del cubo con las letras A, B, C, D y suponer que cada permutación de dichas letras corresponde a una transformación geométrica del cubo. Por ejemplo, la permutación que transforma A en B, B en C, C en A, y deja invariante D corresponde a una rotación espacial de ángulo $2\pi/3$ en torno a la recta D. Se obtiene así una representación de S_4 en un espacio de tres dimensiones.

Grupo de Galois absoluto. El grupo importante para nosotros es el llamado grupo de Galois absoluto, conjunto de las permutaciones de números algebraicos (números complejos solución de ecuaciones polinómicas de coeficientes enteros) que conservan todas las relaciones algebraicas existentes entre ellos. Por ejemplo, un número fraccionario $x = m/n$ (con m y n enteros) verifica la ecuación $nx - m = 0$ y el grupo de Galois sólo puede permutarlo con un número que satisfaga siempre esta relación, es decir, con el propio x: todos los números racionales permanecen inalterados bajo permutaciones del grupo de Galois. Otro ejemplo: el número $1 + i$, solución de la ecuación $x^2 - 2x - 2 = 0$, debe transformarse en otra solución de dicha ecuación. De ahí que se transforme necesariamente en $1 + i$ (él mismo) o en $1 - i$.

El grupo de Galois absoluto es crucial en aritmética porque conserva las propiedades racionales y el estudio de sus representaciones es un importante tema de investigación. Es posible definir representaciones del grupo de Galois asociadas a una curva elíptica y también representaciones del grupo de Galois asociadas a una forma modular. La conjetura STW prevé, como es de esperar, que ambas deben coincidir.

En lo tocante a las formas modulares, la definición exacta de las representaciones de Galois es demasiado técnica para que podamos darla aquí. Di-

gamos simplemente que cada permutación del grupo de Galois da lugar a una «simetría» que puede describirse en forma de matriz (una tabla) 2 × 2; para alguna de estas permutaciones, la suma de dos elementos diagonales de la correspondiente matriz es un coeficiente c_p del desarrollo de la forma modular f.

Dejando aparte la aventura humana que representa, el teorema de Fermat es también un espectacular reflejo del estado actual de la aritmética

En las formas elípticas, son los conjuntos finitos E[*n*] quienes sirven de receptáculo para las representaciones. Recordemos que un punto P(*x*,*y*) está en E[*n*] si *x* e *y* satisfacen la ecuación (1) de la curva y si *n*P = 0, lo cual se traduce en ecuaciones polinómicas entre sus coordenadas. Una permutación del grupo de Galois sobre las coordenadas x e y las transforma en números que satisfacen las mismas relaciones, y por tanto en otro punto P' de E[*n*]. No sabemos reconstruir la curva elíptica E a partir de una sola de estas representaciones. No obstante, un resultado más general, conjeturado por el norteamericano J. Tate y demostrado por Faltings en 1982 dentro de un marco mucho más general, establece que basta el conocimiento de las representaciones E[*k*], E[k^2], E[k^3], etc. (para un número primo *k* conveniente). Esta información puede agruparse en una sola representación del grupo de Galois, llamada k-ádica, en un espacio que engloba la torre de los E[k^r]. La conjetura STW equivale entonces a decir que esta representación coincide con una representación «modular».

Últimos pasos. En los años 1980 se disponía ya de estos ingredientes, pero lograr construir, para toda curva elíptica E, la forma o la representación modular correspondiente —o demostrar que la serie o representación de que se dispone naturalmente son modulares— parecía algo inaccesible. Esto fue precisamente lo que se propuso conseguir Wiles para todas las curvas elípticas E llamadas «semiestables», aquellas para las cuales el nivel N que debe asociárseles no es divisible por ningún cuadrado.

El punto de partida de Wiles fue el número *k* = 3. Gracias a los trabajos del estadounidense J. Langlands,[11] completados por su compatriota J. Tunnell en 1981,[12] sabíamos construir una representación modular que coincide con la representación de Galois sobre E[3]. Pero parecíamos todavía muy lejos del objetivo, porque era preciso establecer la conjetura STW de una representación que coincidiera también con las de E[9], E[27], E[81], etc. Wiles utilizó a tal fin una teoría algebraica de las «deformaciones» de las representaciones de Galois desarrollada por el japonés H. Hida, el estadounidense B. Mazur y el francés J. Tilouine durante el último decenio.[13, 14, 15] Wiles construyó una representación de Galois «universal» sobre un nuevo espacio que contiene tanto las representaciones modulares como las que proceden de las curvas elípticas. En 1993, en una primera versión de su trabajo, Wiles había tratado de realizar un cálculo explícito de la eventual diferencia entre los dos tipos de representaciones; el proyecto, muy complicado, no pudo llevarse a buen puerto y durante unos meses se creyó incluso que había una laguna definitiva en la prueba. En otoño de 1994, con la ayuda de Richard Taylor, enseñante investigador del Instituto Newton de Cambridge (GB), Wiles logró establecer un teorema de estructura más fino para estas deformaciones[16,17] que le permitió no tener que recurrir a construcciones explícitas y completar la demostración de un modo totalmente satisfactorio. Andrew Wiles ha sido galardonado con numerosos premios: además del premio Fermat, el premio Schock Prize in Mathematics de la Academia real de ciencias de Suecia y una medalla especial del congreso internacional de matemáticos de Berlín en 1998 (al tener Wiles un poco más de cuarenta años no podía recibir la famosa medalla Fields). Dejando aparte la aventura humana —que el brillante trabajo de Wiles encarna perfectamente— el teorema de Fermat constituye también un espectacular reflejo del estado actual de la aritmética. La conjetura STW ejemplifica muy bien un conjunto de conjeturas que inervan todas las matemáticas. Está ligada, entre otras cosas, al programa de Langlands que busca relacionar, a través de sus representaciones, objetos geométricos (aquí curvas elípticas) y funciones especiales (como las funciones modulares). Desde 1995, se han hecho importantes progresos en torno a la conjetura STW en general (para otros niveles N) y en otras situaciones ligadas al programa de Langlands. La demostración del teorema de Fermat no ha sido pues el punto final de la teoría de números. ■

CATHERINE GOLDSTEIN es encargada de investigación del CNRS y trabaja en el laboratorio de aritmética y geometría algebraica de la Universidad de París-Sud. Ha investigado en aritmética de las curvas elípticas y actualmente se dedica a la historia de la teoría de números en el Occidente moderno.

(1) J.Itard, *Rev. hist. sc., III*, 21, 1949.
(2) R. Rashed, *Rev. hist. sc., XXXII/3*, 193, 1979.
(3) C. Goldstein, *Un Théorème de Fermat et ses lecteurs*, PUV, Saint-Denis, 1995.
(4) H. Bos, H. Mehrtens I. Schneider, *Social History of Nineteen Century Mathematics*, Birkhauser, 1981.
(5) H. Edwards, *Arch. Hist., Ex. Sc.*, 14, 1975 y 17, 381, 1977.
(6) Y. Hellegouarch, *C.R. Acad. Paris, 273, Série I*, 1194, 1971.
(7) G. Frey, Ann. *Univ. Sarav. Math., 1*, 1, 1986.
(8) J.-P. Serre, *Duke Math., 54*, 179, 1987.
(9) K. Ribet, *Inv. Math., 100*, 431, 1990, *y Ann. fac. sc. Toulouse*, 5, XI, 1990.
(10) J. Oesterlé, *Sém. Bourbaki, 694*, 1987-1988.
(11) R. Langlands, *Ann. Math. Stud.*, 96, Princeton University Press, 1980.
(12) J. Tunnell, *Bull., AMS*, 5, 173, 1981.
(13) H. Hida, *Inv. Math.*, 85, 545, 1986.
(14) B. Mazur, in «Galois Groups over Q», Math., Sc. Res. Inst., 16, Springer Verlag, 1989.
(15) B. Mazur y J. Tilouine, *Pub. IHES, 71*, 1990.
(16) A. Wiles, *Annals of Maths, 141*, 443, 1995.
(17) R. Taylor y A. Wiles, *Annals of Maths, 141*, 553, 1995.

Para más información:

- C. Goldstein, «Le métier des nombres», in *Éléments d'histoire des sciences*, Bordas, 1989.
- A. Weil, *Number Theory: an Approach trough History*, Birkhauser, 1984.
- J.-P. Serre, Cours d'arithmétique, PUF, 1977.
- N. Koblitz, *Introduction to Elliptic Curves and Modular Forms*, GTM, Springer, 1984.
- J. Silverman, *Elliptic Curves*, GTM, Spirnger, 1986.
- S. Lang, «Number Theory III», *Enciclopaedia of mathematical sciences*, Springer, 1991.
- K. Ribet, *Modular elliptic Curves and Fermat's Last Theoremc, 1993*, casete de vídeo (ref. 89NA) disponible en la American Mathematical Society, PO Box 5904, Boston, MA 02206, EEUU.
- Y. Hellegouarch, *Invitation aux mathématiques de Fermat*, Wiles, Masson, 1997.
- S. Singh, *Le Dernier Théorème de Fermat*, Lattès, 1998.
- C. Corrales, C. Andradas, *Cuatrocientos años en torno al último teorema de Fermat*, ed. Complutense, Madrid, 1999.



¿Existen los números infinitos?

Hourya Sinaceur

¿Tiene realidad el infinito, o es una simple «ficción útil» del cálculo, como pensaba Leibniz? La teoría de conjuntos ha definido y cuantificado rigurosamente esta noción. No obstante, dos tendencias filosóficas persisten entre los matemáticos: la que consiste en hacer un uso esencial del infinito y la que prefiere prescindir de él.

Desde siempre, la cuestión del infinito es la que más ha atormentado *la sensibilidad* de los hombres; *la idea* de infinito, la que más ha solicitado y fecundado su inteligencia; *es el concepto* de infinito es el más necesitado de *dilucidación*.» así se expresaba, en 1925, David Hilbert, uno de los más grandes matemáticos de todos los tiempos (véase el artículo de Hourya Sinaceur y Jean-Pierre Bourguignon en el presente número). Hoy en día, no se pueden considerar zanjadas todas las cuestiones, ni siquiera las puramente matemáticas, suscitadas por el infinito. Una de las principales características de este concepto es el hecho de haber atravesado toda la historia de las matemáticas, desde sus orígenes hasta los desarrollos más actuales. En el siglo tercero antes de Cristo, el matemático griego Arquímedes dedicó un largo trabajo a contabilizar los granos de arena que llenan la superficie de la Tierra. Como la totalidad de los granos de arena era inagotable y la enumeración inacabable, Arquímedes llegó a la conclusión de que la sucesión de los números enteros no tenía fin y podía ser prolongada hasta el infinito.

El infinito se presentaba así por su lado aritmético, haciéndose patente al matemático deseoso de contar más allá de los números familiares. Este aspecto aritmético del infinito apareció tardíamente. El infinito, en efecto, se había manifestado muy pronto bajo otras formas. Los pitagóricos (siglo XI a.C.), Zenón de Elea (siglo V a.C.), Eudoxo de Cnido (principios del siglo IV a.C.) y Euclides, por no citar sino unos pocos nombres emblemáticos, ya lo habían experimentado. El infinito aparecía, por ejemplo, cuando se intentaba asignar un número a cualquier magnitud física o geométrica. Los pitagóricos, al tratar de determinar mediante un número (entero o racional) el cociente entre la longitud de la diagonal y la del lado, supuesto igual a 1, de un cuadrado, descubrieron que estas dos magnitudes eran inconmensurables (lo que, en términos modernos, equivale a afirmar que la raíz cuadrada de 2 es irracional).

> Aristóteles admitía la necesidad de pensar el infinito, pero negaba a esta noción toda existencia física o matemática

El infinito aparecía también al pensar que una magnitud continua es divisible hasta el infinito. Así ocurría en las paradojas, como la de la liebre y la tortuga, con las que Zenón pretendía demostrar que el movimiento era imposible: para alcanzar un punto, un móvil tiene que recorrer primero la mitad de la distancia, pero antes la mitad de la mitad y así sucesivamente dividiendo el camino hasta el infinito. Otra ilustración de infinito la dio Arquímedes al mostrar cómo medir un arco de parábola aproximándolo en infinitos pasos mediante arcos de parábola cada vez más pequeños.

El callejón sin salida de los griegos. Así, los matemáticos griegos tropezaban con el infinito al tratar de contar sin fin, de *medir* magnitudes, como la diagonal del cuadrado o el arco de parábola, y de *definir* las propiedades del continuo. Para Aristóteles, primer teórico del infinito y del continuo, el infinito era «lo que no se deja recorrer y carece de límite». Al carecer de límite, no puede ser «determinado» y no existe en sí mismo. En efecto, si una cosa fuera infinita, habría que decir también que sus partes son infinitas. Pero entonces habría que reconocer que un infinito —el del todo— es mayor que otro —el de una de sus partes—.

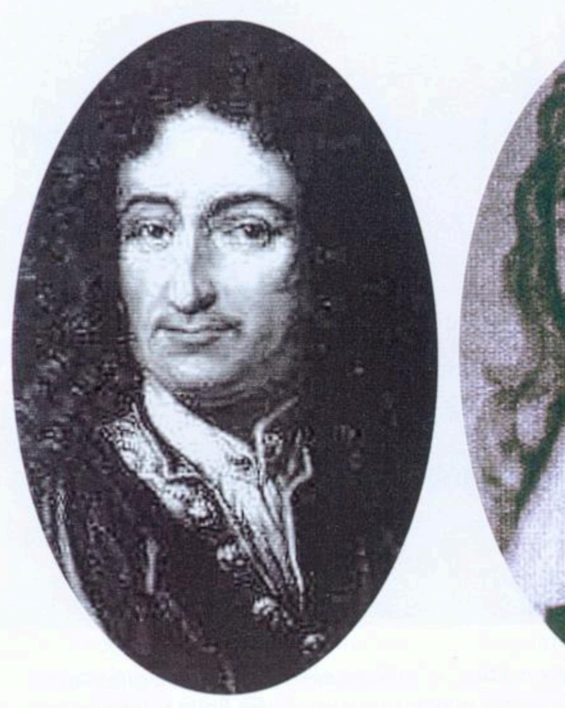
Aparecen aquí los dos principios que impidieron a los griegos concebir un infinito en sí o, como dice Aristóteles, un infinito «en acto»: el axioma, enunciado por Euclides, de que el todo es mayor que cualquiera de sus partes, y la tesis de que no pueden existir infinitos mayores que otros. Por ello, la visión «positiva» de Arquímedes, según la cual la idea del infinito es geométricamente demostrable y está físicamente realizada en el número de granos de arena existentes en la Tierra, no resistió al análisis de Aristóteles, que concebía el infinito como una pura negación de lo finito.

Aristóteles admitía la necesidad de pensar el infinito, pero le negaba toda existencia física o matemática. Para él, el matemático tiene necesidad de considerar magnitudes mayores, o más pequeñas, que toda magnitud dada, pero no totalidades infinitas *en acto*, determinadas aunque no limitadas. El infini-

to matemático está indudablemente relacionado con la categoría de *cantidad*, pero sólo como infinito potencial, cantidad que puede hacerse más grande o más pequeña sin que dicho devenir llegue a transformarse en ser. Esta victoria conceptual del infinito *potencial* sobre el infinito actual ha recorrido los siglos y ha llegado hasta nosotros pese al retroceso definitivo que sufrió en el siglo XIX con la teoría de conjuntos infinitos elaborada por el alemán Georg Cantor (1829-1920).

Matemáticas y teología. No es posible hacer una lista exhaustiva de quienes, en la Edad Media, discutieron o modificaron la concepción aristotélica, referencia universal de filósofos, teólogos y científicos: Epicuro, Papo, Proclo, Filopon, Al Kindi, Al Nayrizi, Thabit ibn Qurra, Avicena, etc.[1] Unos pensaban que el infinito no podía ser objeto de inteligencia o de medida; como decía el filósofo griego Proclo (412-485), el infinito no podía ser admitido por sí mismo sino sólo «con vistas a lo finito». Otros, en cambio, eran infinitistas convencidos y aceptaban que un infinito pudiera ser mayor que otro: el conjunto de los números enteros positivos, por ejemplo, contenía muchos más elementos que el de los enteros pares. El científico árabe Thabit ibn Qurra (836-901), que sostenía esto, observó también que el conjunto de los enteros pares y el de los enteros impares constituían un caso de igualdad de infinitos. Desde el punto de vista matemático, las cuestiones debatidas se reducían a las siguientes: ¿hay un solo infinito o varios?; en caso de que haya varios, ¿cómo distinguirlos y compararlos? ¿Cuándo es posible declarar iguales dos infinitos? ¿Es finita o infinita una parte de infinito? En todas estas cuestiones no dejaron de florecer las paradojas. Dichas paradojas lastrarían durante mucho tiempo la matematización del infinito, mientras que el tema de la infinitud divina introduciría, desde la Edad Media, la concepción teológica de un infinito *cualitativo* como modo de ser (en acto) de un Dios perfecto y omnipotente. La perfección del Ser supremo se oponía a la idea aristotélica de un infinito indeterminado y potencial. Pero el hecho de que esta perfección fuera cualitativa dobló la oposición tradicional cantidad/cualidad con una separación entre el infinito ontoteológico y el infinito matemático. Así, B. Spinoza (1632-1677) opuso todavía el «verdadero infinito», el de la sustancia indivisible, al «falso infinito», el infinito según el número, objeto imaginario. En lo que a las matemáticas se refiere, el nacimiento de la física galileana relanzó el interés por el tema. La necesidad de definir los conceptos de velocidad instantánea y de aceleración, así como de formular las leyes del movimiento y de generalizar el concepto de curva, desembocó en la elaboración de los conceptos de función y diferencial.

despreciable ante un infinito de orden superior (por ejemplo, x es despreciable ante x^2 cuando x tiende a infinito); asimismo, un infinitésimo es despreciable ante otro infinitoésimo de orden inferior (por ejemplo, $1/x^2$ es despreciable ante $1/x$ cuando x tiende a infinito). Una jerarquía *operatoria*, basada en la rapidez de crecimiento de las *funciones* que representan a los infinitos, permitió dar respuesta a algunas de las cuestiones planteadas más arriba. Pero no todas las dificultades estaban vencidas. Galileo había considerado la colección de los enteros positivos (1, 2, 3, ...) y la de sus cuadrados (1, 4, 9, ...). Constatando que todo entero tiene un cuadrado, y recíprocamente que todo cuadrado procede de un entero positivo, había llegado a la conclusión de que las relaciones de igualdad y desigualdad no son válidas en el infinito. Leibniz, que había elaborado reglas de igualdad y desigualdad para el infinito, no podía aceptar esta conclusión. Su *Análisis de los infinitos* mostraba precisamente «una nueva manera de sumar, restar, multiplicar, dividir y extraer raíces propia de las cantidades incomparables, es decir, de las que son infinitamente grandes o infinitamente pequeñas con respecto a otras».[2] Pero Leibniz no puso en entredicho la validez del axioma euclídeo según el cual el todo es mayor que las





Wilhelm Gotfried Leibniz e Isaac Newton
El cálculo infinitesimal, la nueva «ciencia del infinito» inventada simultáneamente por I. Newton (1642-1727) y W. G. Leibniz (1646-1716).

La paradoja de Galileo. El cálculo infinitesimal, la nueva «ciencia del infinito» inventada simultáneamente por Isaac Newton (1642-1727) y Wilhelm Gottfried Leibniz (1646-1716), introdujo los «elementos infinitesimales», o infinitésimos, que representan cantidades infinitamente pequeñas. Se distinguían allí distintos órdenes de infinito y se establecían las reglas que permitían compararlos entre sí y con lo finito: un infinitésimo sumado o restado a una cantidad finita es despreciable porque es «incomparablemente» más pequeño que ella; no se cambia el orden de un infinito añadiéndole una cantidad finita (por ejemplo x es del mismo orden que $x + 100$ cuando x tiende a infinito); un infinito de orden inferior es



Desde el punto de vista matemático, el infinito suscita múltiples cuestiones. Por ejemplo, ¿puede ser un infinito mayor que otros? ¿Es finita o infinita una parte de un infinito? ¿Cómo comparar la infinidad de los puntos de una recta con la infinidad de los puntos de un plano? Al tratar de responder a estas preguntas, los científicos topaban a menudo con paradojas. Sólo a fines del siglo XIX, con la aparición de la teoría de conjuntos, se pudo precisar satisfactoriamente la noción de infinito. Pero no por ello quedó cerrado el debate sobre el papel que el infinito debe desempeñar en matemáticas.

partes; por el contario, trató de demostrarlo. Se preguntó, pues, por la posibilidad de que la paradoja de Galileo se debiera a la consideración de estas colecciones como todos acabados.

¿Realidad o ficción? Leibniz concebía lo infinitamente pequeño y lo infinitamente grande como cantidades auxiliares destinadas a facilitar cálculos cuyos resultados finales se expresaban en términos de cantidades finitas, pero que en sí mismas carecían de consistencia. En el fondo seguía en la óptica, definida por Proclo a partir de Aristóteles, del infinito considerado «con vistas a lo finito». Un infinitésimo era una cantidad evanescente que en ocasiones no era nada (comparada con una cantidad finita) y en ocasiones era algo (comparada con un infinitésimo de orden superior); un infinito era una cantidad asintótica que nunca alcanzaba el límite infinito hacia el que tendía. En el siglo XVIII, estas cantidades evanescentes y asintóticas eran la nueva versión, formada con la ayuda del concepto de función, de la concepción aristotélica de un infinito en potencia, nunca en *acto*. En tales condiciones era difícil poner en pie de igualdad las cantidades finitas, a las que se concedía realidad tanto en el mundo como en las matemáticas, y a las cantidades infinitas, a las que se negaba «realidad» matemática, ya que no les correspondía nada sustancial. Para Leibniz, estas últimas eran meras «ficciones» útiles para el cálculo, que abreviaban el razonamiento, análogas, por ejemplo a $i = \sqrt{-1}$. «*El cálculo infinitesimal es útil cuando se trata de aplicar las matemáticas a la física, pero no pretendo con él dar cuenta de la naturaleza de las cosas*», escribió.[3] Así pues, Leibniz, como matemático, concedía al infinito un estatuto que legitimaba su manejo sin que ello supusiera admitir la existencia del infinito actual. No obstante, como metafísico, Leibniz no vacilaba en reformar la noción de sustancia a fin de admitir el infinito en acto. «*Soy tan partidario del infinito actual*, escribió Leibniz en una carta a Foucher, *que en vez de admitir que la naturaleza lo aborrece, considero que lo exhibe universalmente para mostrar mejor las perfecciones de su Autor*». ¿Cómo conciliar esta audacia metafísica con la prudencia matemática? Por medio de distinciones conceptuales. Leibniz sostuvo que incluso en el mundo físico las partes no necesariamente existen en forma de elementos separados, como los famosos granos de arena mencionados por Arquímedes; en un infinito en acto no hay partes en acto. Por otra parte, la división del continuo no implica su composición a partir de elementos atómicos. En último análisis, «*no hay números infinitos, ni líneas u otras cantidades infinitas, si se las toma por verdaderos todos*».[4] En suma, el infinito en acto existe pero no es numerable. De ahí dedujo que las operaciones aritméticas se aplican sólo al infinito potencial. Pese a su metafísica infinitista, el Leibniz matemático, en definitiva, se mantuvo fiel a la tradición definida por Aristóteles.

La historia moderna del infinito matemático comienza con Bernardo Bolzano (1781-1848). Bajo de título de *Las paradojas del infinito*, este matemático checo de origen italiano, que también fue físico, lógico, filósofo y teólogo, escribió una defensa e ilustración del infinito actual basada en la idea de que las supuestas paradojas que recorrieron los siglos desde Zenón de Elea no resistían un análisis consecuente. Su objetivo principal era situar el «verdadero» infinito en el campo del cálculo y de la cantidad y no en Dios, así como convertir el concepto matemático en el fundamento de sus homólogos físico y metafísico. A partir de entonces la teo-

logía quedaba subordinada a la matemática del infinito en *acto*. Y en ésta prevalecía, por supuesto, el punto de vista cuantitativo. Para Bolzano, Dios es infinito porque le atribuimos unas capacidades de magnitud infinita. Indudablemente, Bolzano no fue el primero en afirmar la existencia positiva del infinito *actual*, ni el primero en darse cuenta de la posibilidad de que existieran varios infinitos desiguales, como tampoco el primero en vincular la igualdad entre dos infinitos a la posibilidad de establecer una correspondencia biunívoca entre sus elementos. Pero fue el primero en tratar de construir un concepto puramente matemático y un cálculo sistemático del infinito *actual*. La construcción de Bolzano se basa en un paralelismo bastante estrecho entre lo finito y lo infinito. El estatuto lógico es el mismo: el infinito *actual* es tan poco contradictorio como los conceptos familiares de número entero o de fracción. El estatuto matemático también es el mismo: hay conjunto infinitos en *acto* que nada impide lógicamente concebir como todos acabados. Así, el conjunto de los números enteros, una recta infinita o un segmento comprenden una infinidad de elementos que conceptualmente están perfectamente determinados. No es necesario enumerar todos los elementos de un conjunto para concebir su existencia. Basta *caracterizar* el conjunto mediante una o varias propiedades: una relación de recurrencia simple define la sucesión de los enteros, los puntos determinan un segmento o una recta, etc. También es el mismo, por último, el estatuto ontológico: tal como sostenía Leibniz, el infinito *actual* está realizado en todas las cosas existentes.



Bernardo Bolzano hizo dar a la noción de infinito un paso importante hacia la modernidad. Este científico checo de origen italiano pretendía situar el «verdadero» infinito no en Dios sin el el cálculo y la cantidad. Para Bolzano, el infinito existe en sí mismo y no es fruto de la imaginación. Este científico fue el primero en tratar de construir un concepto puramente matemático del infinito y un cálculo sistemático del mismo. No obstante, su construcción seguía demasiado de cerca los resultados válidos en el caso finito y no permitía caracterizar o comparar con precisión los infinitos. (Foto Seuil)

Igualdad de infinitos. El cálculo, por su parte, se basa en las definiciones siguientes: un infinito es algo «mayor que un número cualquiera de unidades», es decir, mayor que n para todo entero positivo n; un infinitésimo es algo cuya multiplicación por un entero n cualquiera es menor que la unidad. Estas definiciones contradicen el axioma de Arquímedes según el cual dadas dos magnitudes desiguales existe siempre un múltiplo de la menor superior a la mayor. Bolzano no discutió explícitamente este axioma, como tampoco refutó explícitamente el axioma bimilenario del todo y la parte considerando que los conjuntos infinitos están precisamente caracterizados por la posibilidad de ponerlos en correspondencia biunívoca con uno de sus subconjuntos. Lo que unos veían como una paradoja le parecía a él lo *propio* del infinito. No era de extrañar, pues, que se pudiera poner en correspondencia biunívoca los puntos del lado de un cuadrado con los de su diagonal, o el conjunto de los enteros con el de sus cuadrados. Desde el punto de vista del conjunto de sus elementos, lado del cuadrado y diagonal o sucesión de enteros o sucesión de cuadrados representaban el mismo *infinito*. Pero Bolzano no sacó todas las consecuencias de esta observación. No definió, como hizo más tarde Cantor, la igualdad de dos infinitos por la posibilidad de encontrar una biyección de uno en otro, esto es, una correspondencia biunívoca entre los elementos de los conjuntos considerados.

Al tratar de establecer una aritmética del infinito, Bolzano, en efecto, cayó en la trampa de lo finito y definió la igualdad de los dos conjuntos infinitos E y F por su identidad y su desigualdad por la inclusión estricta de uno en otro. Bolzano dio como ejemplo el conjunto de los puntos del segmento [0,5] de la recta, que al estar contenido en el segmento [0,12] es «más pequeño» que él. Pero, ¿qué ocurre si al mismo tiempo E es biyectivamente aplicable en F, como ocurre en este ejemplo al definir la aplicación $x \rightarrow 12x/5$. ¿Se mantendrá que E y F representan el mismo infinito? ¿O se volverá al axioma secular que sienta la desigualdad del todo y la parte? ¿Y de qué sirve afirmar que hay una infinidad de infinitos si no se asignan *números* a todos ellos?

Para Bolzano, Dios sólo es infinito porque le atribuimos unas capacidades de magnitud infinita

Después de lo finito, ¿lo transfinito? Como es sabido, el problema fue resuelto por Georg Cantor con la teoría de conjuntos, para la cual elaboró una aritmética específica. El acto decisivo consistió en afirmar que «después de lo finito, hay un *transfinito*, una escala ilimitada de modos determinados que por naturaleza son infinitos y que, no obstante, pueden ser precisados, como en el caso finito, mediante números determinados, bien definidos y distinguibles unos de otros». Era la primera vez en matemáticas que se hablaba de *números* infinitos. O al menos de números transfinitos, es decir, infinitamente grandes, pues Cantor no admitía la existencia de los infinitésimos y hubo que esperar al «análisis no estándar», en 1966, para reconocerlos, por fin, como entidades bien definidas. Cantor se vio llevado a elaborar la teoría de conjuntos al estudiar los puntos de discontinuidad de las funciones representables por series trigonométricas (suma de una infinidad de términos de la forma c_n sen $nx + d_n \cos nx$). En el marco de estos estudios, introdujo la definición de los números reales como límites de sucesiones de números racionales y formuló el axioma de la correspondencia biunívoca entre los números reales y la recta del plano (estos conjuntos son, respectivamente, las representaciones numérica y

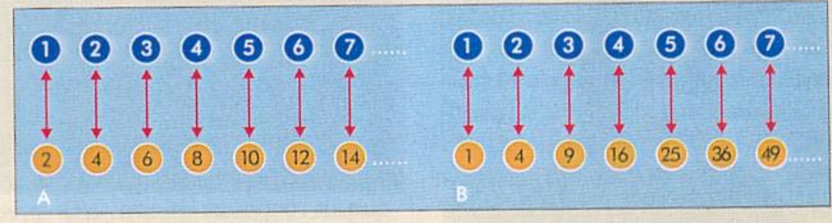
geométrica del «continuo lineal»). En 1883 descubrió los «ordinales transfinitos» y propuso entonces de un modo totalmente explícito una generalización al transfinito de la noción de número entero finito.

Cantor empezó por establecer una clara distinción entre número *cardinal* y número *ordinal*. La primera noción está vinculada a la operación de contar los elementos de un conjunto, que es indiferente al orden de éstos, y la segunda a la de enumerar, que establece un orden entre los elementos. Estas dos nociones se confunden en el caso finito: se enumeren como se enumeren los elementos de una colección finita, el último elemento enumerado, digamos que el n-ésimo, marca al mismo tiempo el número cardinal de la colección, n. Pero no ocurre lo mismo en las colecciones infinitas. Tomemos ejemplos simples. Sean los dos conjuntos $E = \{1,2,3,4,...\}$ y $F = \{2,3,4,....1\}$, que contienen los mismos elementos ordenados de modo distinto. E y F tienen el mismo cardinal, pero ordinales diferentes. Por contra, los conjuntos finitos $E = \{1,2,3\}$ y $F = \{3,2,1\}$, que tienen el mismo cardinal y cuyos elementos no están ordenados igual, tienen también el mismo ordinal. Lo que es idéntico en ellos no es el orden sino la *enumeración*: se empieza por un primero y se termina por un tercero independientemente de los elementos que se hayan colocado como primero, segundo y tercero.

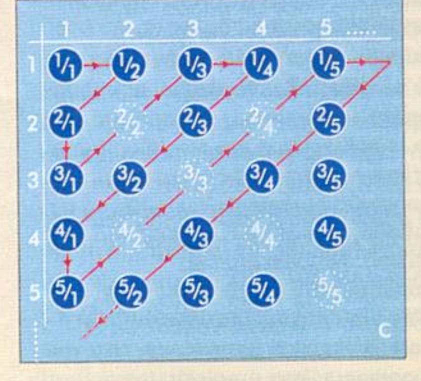
Precisemos en primer lugar la noción de cardinal transfinito, más simple porque prescinde de la estructura de orden de los conjuntos considerados. Cantor estableció que dos conjuntos infinitos tienen la misma *potencia*, o el mismo *cardinal*, si existe una biyección de uno en otro (véase el recuadro «Cardinales transfinitos»). Así ocurre con el conjunto de los enteros y el de sus cuadrados, con el conjunto de los números pares y el de los enteros, con el conjunto de los puntos de dos segmentos cualesquiera, etc. Se dice que los conjuntos entre los que se puede establecer una aplicación biyectiva son «equipotentes». Esta definición conviene también para los conjuntos finitos, en cuyo caso la equipotencia coincide con la igualdad. Con ello se vencían por primera vez las dificultades que impedían el establecimiento de una aritmética transfinita, la principal de las cuales es que nadie antes que Cantor había tratado de generalizar al

Cardinales transfinitos





El matemático alemán Georg Cantor es el padre de la teoría de conjuntos. Esta teoría ha permitido definir y cuantificar el infinito de manera precisa y rigurosa, gracias sobre todo a la noción de cardinal transfinito. El cardinal de un conjunto indica el número de elementos que éste contiene y permite extender la noción de número a los números infinitos. Dicho con mayor rigor, dos conjuntos (finitos o infinitos) son equipotentes o tienen el mismo cardinal si existe una correspondencia biunívoca entre ellos. Por ejemplo, el conjunto de los números enteros positivos tiene el mismo cardinal que el conjunto de los enteros pares (A) o que el conjunto de los cuadrados enteros (B). En este sentido, dichos conjuntos tienen «el mismo número» de elementos, contrariamente a lo que parece desprenderse de la intuición. Asimismo, el conjunto de los números racionales es equipotente al conjunto de los enteros; en efecto, se pueden enumerar todos los números racionales, sin excepción, en el orden 1/1, 1/2, 2/1, 3/1, 1/3, 1/4, 2/3, etc. La sucesión así obtenida define una biyección entre el conjunto de los enteros positivos y el conjunto de los racionales positivos. Esta enumeración exhaustiva se representa en C (donde se omiten los números racionales que se repiten). (Foto A: J.-L. Charmet)

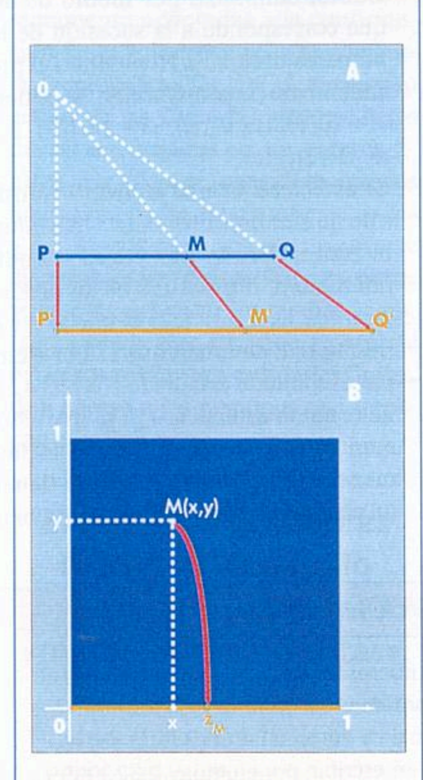


dominio transfinito la noción de número entero finito. La escala de los cardinales transfinitos comienza con el cardinal del conjunto N de los números naturales, que es *numerable*. Este cardinal se representa por $\aleph_0$ (el símbolo $\aleph$ —alef— es la primera letra del alfabeto hebreo). Otros muchos conjuntos infinitos son equipotentes a N. Por ejemplo, Q_+, el conjunto de los números racionales positivos, tiene el mismo cardinal que N, ya que se pueden enumerar todos los racionales, sin excepción, en el orden 1/1, 1/2, 2/1, 3/1, 1/3, 1/4, 2/3, 3/2, 4/1, 5/1, 1/5, 1/6, 2/5, 3/4, 4/3, 5/2, 6/1, etc. (véase el recuadro «Cardinales transfinitos»). La sucesión así obtenida define una biyección de N en Q_+. Asimismo, Cantor demostró en 1874 que el conjunto de los números algebraicos, es decir, de los números reales que son raíces de un polinomio de coeficientes enteros, también es equipotente a N.

¿Cuántos cardinales? En cambio, demostró que el conjunto de los números reales R no es equipotente a N. Se dice que R tiene la potencia del continuo, representada por c. La demostración de que R tiene una potencia estrictamente mayor que la de N se hace por el absurdo, recurriendo a un método inventado por Cantor, el famoso procedimiento de la diagonal (véase el recuadro «El método de la diagonal»). Cantor descubrió luego que existe una biyección entre el conjunto de los puntos de un cuadrado cuyo lado es el intervalo [0,1] y este mismo intervalo (véase recuadro «Equipotencia»). Este resultado le sorprendió hasta el punto de escribir lo siguiente a Dedekind (1831-1916), con el que mantenía una correspondencia periódica: «Lo veo pero no lo puedo creer». En 1891, Cantor, por el método diagonal, demostró que para todo conjunto infinito E, el conjunto de las partes de E tiene una potencia estrictamente mayor que la de E. Dicho resultado es absolutamente fundamental: significa que no hay sólo dos cardinales transfinitos: el numerable $\aleph_0$ y el continuo c, sino una infinidad. Se obtiene, por tanto, una generalización de los números cardinales finitos, aunque con reglas aritméticas que difieren en parte de las que rigen para lo finito.[5] En efecto, definiendo la suma de los cardinales de dos conjuntos infinitos disjuntos E y F como el cardinal de la reunión de E y F, se obtiene, por ejemplo: $\aleph_0 + n = \aleph_0 + \aleph_0 = \aleph_0 + \aleph_0 ... + \aleph_0 + ... = \aleph_0$, y para todo cardinal transfinito c, $c + n = c + \aleph_0$. Análogamente, para la multiplicación, la biyección mencionada antes entre los puntos del cuadrado de lado 1 y los del segmento de longitud 1 se traduce en la operación $c \bullet c = c$. No obstante, pese a estas diferencias, la asociatividad, la conmutatividad y la distributividad siguen siendo válidas para la suma y la multiplicación de cardinales transfinitos.

En este punto, surge un problema: ¿es posible comparar siempre dos cardinales cualesquiera? Dados dos conjuntos E y F, es posible que se logre demostrar la existencia de una biyección entre ambos, y que, por tanto, $\mathrm{Card}(E) = \mathrm{Card}(F)$. También puede haberse demostrado que E es equipotente a una parte de F, en cuyo caso $\mathrm{Card}(E) = \mathrm{Card}(F)$, o que F es equipotente a una parte de E y que $\mathrm{Card}(F) \leq \mathrm{Card}(E)$. Pero para que la comparación fuera siempre posible, sería necesario excluir la eventualidad de un cuarto caso. Sólo entonces se podrían disponer todos los cardinales en una sucesión análoga a la de los números enteros 1,2,3,..., esto es, en un conjunto totalmente ordenado y bien ordenado por la relación «mayor que».

Equipotencia



Dos segmentos PQ y P'Q' de longitud diferente tienen el mismo cardinal; en términos menos precisos, se puede decir que contienen «el mismo número» de puntos. En efecto, se puede hacer corresponder a cada punto M del primer segmento un punto y uno solo M' del segundo y recíprocamente. Una tal biyección se representa en A. Más sorprendente todavía es el hecho de que el conjunto de los puntos contenidos en un cuadrado tiene el mismo cardinal que el conjunto de los puntos de su lado. Efectivamente, se puede hacer corresponder a todo punto M del cuadrado de lado 1 un punto único del intervalo [0,1]: dado M por sus coordenadas x e y, cuya representación decimal es $x = 0{,}x_1x_2x_3...$ e $y = 0{,}y_1y_2y_3...$, se puede asociar a M el punto del intervalo [0,1] cuya escritura decimal es $z = 0{,}x_1{,}y_1{,}x_2y_2x_3y_3$. Inversamente, es posible escribir todo punto z del intervalo [0,1] de manera que le corresponda un par único (x,y), es decir, un punto único M del cuadrado. El cuadrado y su lado son, pues, equipotentes (B).

Los tipos de orden de los conjuntos bien ordenados constituyen los números ordinales

Hénos pues ante el problema de la relación de orden. Un conjunto ordenado es un conjunto provisto de una relación de orden, representada por ≤. Se dice que dos conjuntos ordenados tienen el mismo tipo de orden si son isomorfos, es decir, si hay una biyección f de uno en el otro que preserva el orden: $f(x) \leq f(y)$ cuando $x \leq y$. Por ejemplo, los conjuntos ordenados $A = \{1,2,3,...\}$ y $B = \{...,3,2,1\}$ no tienen el mismo tipo de orden; no es posible hallar una biyección f que aplique A en B de tal modo que $f(1)$ sea el primer elemento de B, pues B carece de primer elemento.

Se dice que un conjunto está *bien ordenado* si toda parte no vacía de E posee un primer elemento. Un conjunto finito está necesariamente bien orde-

nado. La sucesión infinita de los enteros naturales es un conjunto bien ordenado; lo mismo cabe decir de toda sucesión indexada por N. Los tipos de orden de los conjuntos bien ordenados constituyen los números ordinales. Notemos que hay tipos de orden que no son ordinales: por el ejemplo, el tipo de orden de $B = \{...,3,2,1\}$ no es un ordinal porque B no está bien ordenado.

Una hipótesis indecidible. Cantor consideró obvio que había que dar prioridad a los conjuntos bien ordenados, ya que lo que trataba de hacer era reproducir la situación de los enteros. Tenía razón, siempre que lograra demostrar que todo conjunto puede ser bien ordenado. Es lo que hizo en 1904 el alemán Ernst Zermelo. Por medio de su teorema de «buen orden» se puede transferir a los conjuntos infinitos algunas de las propiedades de los números finitos, especialmente la propiedad de inducción en la que se basa el razonamiento por recurrencia.

Tal como acabamos de presentarlo, un ordinal caracteriza una clase de conjuntos bien ordenados isomorfos. Mediante un procedimiento usual en matemáticas, se puede también definir un ordinal como un representante particular de la clase que caracteriza. Se construyen los ordinales finitos sentando $0 = \phi$ (el conjunto vacío), $1 = \{0\}$, $2 = \{0,1\}$ ..., $n + 1 = \{0, 1, 2,..., n\}$, y así sucesivamente. El primer ordinal transfinito es el ordinal límite $\{0, 1, 2,...\}$, que Cantor simbolizó por medio de ω y que corresponde a la sucesión de los números naturales. Repitiendo el procedimiento de construcción, se obtiene $\omega + 1 = \{0, 1, 2,..., \omega\}$, $\omega + 2 = \{0, 1, 2,..., \omega, \omega + 1\}$, etc.

Se establece así una jerarquía transfinita de números ordinales. Un conjunto bien ordenado cualquiera χ tendrá por número ordinal uno de los ordinales $0, 1, 2,..., \omega, \omega + 1,...,$ el que tenga el mismo tipo de orden que él. Por ejemplo, el conjunto ordenado $F = \{2,3,4...1\}$ tiene como ordinal $\omega + 1$, ya que estos conjuntos se enumeran de la misma manera (en particular, ambos tienen un último elemento). Dado que, como hemos vista más arriba, F tiene el mismo cardinal que el conjunto $E = \{1,2,3,4,...\}$, cuyo ordinal es ω, se deduce que a un mismo cardinal le corresponden varios ordinales.

Es tiempo ya de volver a la noción de cardinal. Hemos visto que, intuitivamente, un cardinal de un conjunto cuenta el número de sus elementos, pero es posible dar una definición más precisa: el cardinal de E es el más pequeño ordinal equipotente a E. Esta construcción rigurosa garantiza la posibilidad de ordenar todos los cardinales en una sucesión bien ordenada: $\aleph_0$, $\aleph_1, ..., \aleph_n, ..., \aleph_\omega, \aleph_{\omega+1}, ..., \aleph_{\omega 2}, \aleph_{\omega 2+1}, ...$

Cantor intentó averiguar si el cardinal c del continuo era el sucesor inmediato del cardinal de los conjuntos numerables, en otras palabras, si $c = \aleph_1$. La respuesta afirmativa a esta cuestión constituye la famosa «hipótesis del continuo», que Cantor formuló por primera vez en 1878 y que trató en vano de demostrar hasta el fin de sus días. Hilbert la inscribió como problema en el encabezamiento de su famosa lista propuesta al Congreso internacional de matemáticos de 1900. La respuesta vendría en dos tiempos. En 1938, el físico austriaco Kurt Gödel demostró la compatibilidad de esta hipótesis con los axiomas habituales de la teoría de conjuntos, los axiomas de Zermelo-Fraenkel. Más tarde, en 1963, el norteamericano Paul Cohen demostró su independencia con respecto a dichos axiomas. La conjunción de ambos resultados significa que la hipótesis del continuo es *indescifrable* en la axiomática de Zermelo-Fraenkel: no es posible demostrarla ni refutarla sobre la base exclusiva de dichos axiomas.

EL MÉTODO DE LA DIAGONAL

Para demostrar que el conjunto de los números reales no es equipotente al conjunto de los enteros naturales, se considera el conjunto E de los números reales del intervalo [0,1]. Cada elemento e de E posee un desarrollo decimal; dicho desarrollo es único si se conviene en escribir, por ejemplo, 0,5239999... en vez de 0,524000... Razonemos por el absurdo. Si E fuera numerable, se podrían enumerar todos sus elementos de la manera siguiente:

$e_1 = 0,a_{11}a_{12}a_{13}a_{14}...$
$e_2 = 0,a_{21}a_{22}a_{23}a_{24}...$
$e_3 = 0,a_{31}a_{32}a_{33}a_{34}...$
$e_4 = 0,a_{41}a_{42}a_{43}a_{44}...$

donde cada a_{ij} es una de las cifras 0,1,2,...9. Se comprueba fácilmente la posibilidad de construir un elemento de E que no figure en la enumeración. Basta considerar el número decimal $b = 0,b_1b_2b_3...$, definido a partir de la diagonal de nuestra enumeración por medio de $b_i = 2$ si $a_{ii} = 1$ y $b_i = 1$ si $a_{ii} \neq 1$. Así construido, b difiere de ei por su i-ésima cifra para todo i. En consecuencia, el número b no forma parte de la lista, contrariamente a la hipótesis de partida, lo que demuestra que E no es numerable.

Señalemos que mediante un razonamiento análogo Gödel demostró en 1931 sus famosos teoremas de incompletitud de la aritmética.

Cantor trató en vano, hasta el fin de sus días, de demostrar la famosa «hipótesis del continuo» que había formulado en 1878

Disputa en el paraíso. Con los trabajos de Cantor, la disputa del infinito renació en un momento en que parecía ya sin objeto. Desde sus inicios, la teoría de los números transfinitos había suscitado grandes resistencias. Su adversario más encarnizado era el alemán Leopold Kronecker, que trató de impedir la publicación de los primeros artículos de Cantor, aquellos en que demuestra que el conjunto de los números reales no es numerable y que el conjunto de los puntos de un cuadrado es equipotente al conjunto de los puntos de su lado. Pero Dedekind hizo valer su influencia en favor de Cantor. Más tarde, Hilbert defendió encarnizadamente el «paraíso» de la creación cantoriana, que consideraba como «el producto más puro del ingenio matemático». Pero las paradojas descubiertas en la teoría de conjuntos, la primera por el propio Cantor, socavaban el edificio. Los matemáticos se dividieron en dos bandos: el de los cantorianos que, como Hilbert, hicieron todo lo posible para que «nadie nos expul-

se del paraíso» y el de los demás que, como Kronecker, Henri Poincaré y más tarde Brouwer y Hermann Weyl, se mantuvieron fieles, como antaño Aristóteles, al infinito potencial y numerable sin saltar a las abstracciones de lo transfinito. Estos últimos consideraban que sólo los números enteros son objeto de una intuición indiscutible, que vienen dados por una sucesión de longitud no acotada y no como *conjunto* o totalidad acabada. Tuvieron lugar ásperas discusiones a propósito de un equivalente del teorema del buen orden, el llamado «axioma de elección». Este axioma, enunciado por Zermelo en 1904, estipula que para toda familia, aunque sea infinita, de conjuntos no vacíos, existe una función (no especificada) que a cada conjunto le asocia uno de sus elementos, permitiendo construir así un nuevo conjunto no vacío. El asunto era importante, pues no tardó en quedar claro que muchos resultados (por ejemplo, el que afirma que todo espacio vectorial tiene una base), pertenecientes a otras ramas de las matemáticas, no pueden ser rigurosamente demostrados sin dicho axioma. La historia del axioma de elección corre paralela a la de la hipótesis del continuo. Gödel demostró en 1938 su compatibilidad con los demás axiomas de la teoría de conjuntos, eliminando definitivamente el temor de que su utilización pudiera llevar a contradicciones. Cohen demostró además su independencia, lo cual permitió concluir que se trata de una proposición indescifrable, adjuntable a voluntad a los demás axiomas de la teoría de conjuntos.

Los resultados de indecidibilidad acabaron con las controversias al demostrar la legitimidad lógica de opciones contradictorias. No es de extrañar, pues, que desde fines del siglo pasado florecieran unas matemáticas conjuntistas cada vez más abstractas (topología general, teoría de la integración, análisis funcional, análisis no estándar, etc.) que no vacilan en recurrir al axioma de elección o a sus equivalentes, al lado de matemáticas constructivas, que exploran la potencia de los métodos restringidos a lo finito y lo numerable. Hay que reconocer que la utilización cada vez menos auxiliar y cada vez mas sistemática de los ordenadores impulsa a los matemáticos a buscar réplicas algorítmicas de las disciplinas tradicionales del infinito y del continuo (geometría, análisis y topología), directamente adaptadas a la confirmación finita y discreta de la herramienta tecnológica. De ahí el desarrollo actual de las llamadas «matemáticas finitistas», basadas en los enteros finitos. En ellas, se considera que las únicas entidades efectivamente dadas y los únicos procesos efectivamente ejecutables son finitos. Por otra parte, se intenta delimitar los medios (construcciones, reglas, etc.) que, *a partir* de procesos relativos a entidades finitas, dan acceso a las nociones que implican el infinito.

El uso de ordenadores motiva el desarrollo de las llamadas «matemáticas finitistas», basadas en los enteros finitos

Es curioso constatar que el análisis no estándar, el cual, en sus orígenes, tenía una orientación decididamente infinitista y actualista para hacer aceptar la idea de una extensión del conjunto de los números reales por medio de elementos infinitos, se vuelve actualmente hacia las técnicas finitistas. Varios trabajos recientes permiten introducir un modelo finitista de los números reales y del continuo.(6)

Las matemáticas finitistas llevan a preguntarse por la necesidad teórica de asumir toda la escala de los cardinales transfinitos de Cantor. Esta cuestión, ya planteada por el francés Emile Borel, ha sido estudiada actualmente por el lógico norteamericano Solomon Feferman, de la Universidad Stanford. La respuesta es que, para las matemáticas *útiles* (aplicables al mundo físico), no hay ninguna obligación *lógica* de aceptar el infinito actual. A partir de las ideas expuestas por H. Weyl en su monografía *Das Kontinuum* (1918), Feferman demuestra que es posible limitarse a lo infinito numerable en todos los problemas planteados por las aplicaciones del análisis clásico o del análisis funcional moderno. ¿Hay que renunciar por completo, pues, a la noción de número real? No, ya que muchos resultados matemáticos, considerados en toda su generalidad, hacen un uso esencial de lo transfinito.

Así pues, la dualidad finito/infinito sigue trazando en las matemáticas una línea divisoria que los matemáticos redefinen constantemente sin llegar a abolirla. Es relativamente sencillo reconocer la validez de un resultado a partir de hipótesis admitidas, pero lo es mucho menos ponerse de acuerdo acerca de las hipótesis que hay que admitir. Como escribió Henri Lebesgue, «los matemáticos nunca han estado totalmente de acuerdo sobre el conjunto de su ciencia, que supuestamente es el de las verdades evidentes, absolutas, indiscutibles, definitivas. ■

HOURYA SINACEUR es directora de investigación del CNRS, en el que preside la 35ª sección (Pensamiento filosófico -Ciencia de los textos -Creación artística, científica y técnica). Trabaja en el Instituto de historia y filosofía de la ciencia y de la técnica (Universidad París I)

(1) Para una breve exposición de la prehistoria antigua y medieval del infinito, se puede consultar la obra de Tony Lévy, *Figures de l'infini*, Seuil, 1987.
(2) *Leibnizens mathematischen Schriften*, ed. Gerhardt, Berlín, 1848-1863, reimpr. Hildesheim, Olms, 1962, tomo V, p. 259.
(3) Carta a Masson, *Leibnizens Philosophische Schriften*, ed. Gerhard, Berlín, 1875-1890; reimpr. Hildesheim, Olms, 1965, tomo VI, pág. 629.
(4) W.G. Leibniz, *Nuevos ensayos sobre el entendimiento humano*, libro II, cap. XVII.
(5) *Ibid.*, p. 280.
(6) Se pueden encontrar varios ejemplos en *Le labyrinthe du continu*, editado por H. Sinaceur y J.-M. Salanskis, Springer-Verlag, 1992, 4ª parte.

Para Más Información:

- M. Blay, *Les Raisons de l'infini*, Gallimard, 1993.
- F. Burbage y N. Chouchan, *Leibniz et l'infini*, PUF, 1993.
- Cantor-Dedekind, Correspondencia, en J. Cavaillès, *Oeuvres complètes de philosophie des sciences*, Hermann, 1994.
- J. Cohn, *Histoire de l'infini*, trad. francesa, Le Cerf, 1994.
- J. Dauben, *Georg Cantor. His Mathematics and Philosophy of the Infinite*, Cambridge, 1979.
- S. Feferman, «Infinity in mathematics: is Cantor necessary?», *L'infinito nella scienza*, Instituto della Enciclopedia Italiana, 1987, p. 151-210
- T. Lévy, *Figures de l'infini*, Seuil, 1987.
- R. Rashed, *Les mathématiques infinitésimales du IXè au XIè siècle*, vo. II, *Ibn al-Haytham*, Londres, al-Furqan, Islamic Heritage Foundation, 1993.
- L. Schwartz, *analyse I (théorie des ensembles et topologie)*, Hermann, 1991.
- S. Gibilisco ed McGraw-Hill, *En busca del infinito*, Madrid, 1991.
- H. Sinaceur y J.-M. Salanskis (eds.), *Le Labyrinthe du continu*, Springer-Verlag, 1992.

David Hilbert, rigor y simplicidad

Hourya Sinaceur y Jean-Pierre Bourguignon

Las investigaciones del matemático alemán David Hilbert versaron sobre casi todas las ramas de las matemáticas, en particular la teoría de números. Fue uno de los fundadores del método axiomático y desempeñó también un papel determinante en la emergencia de una reflexión sobre los fundamentos de su disciplina.

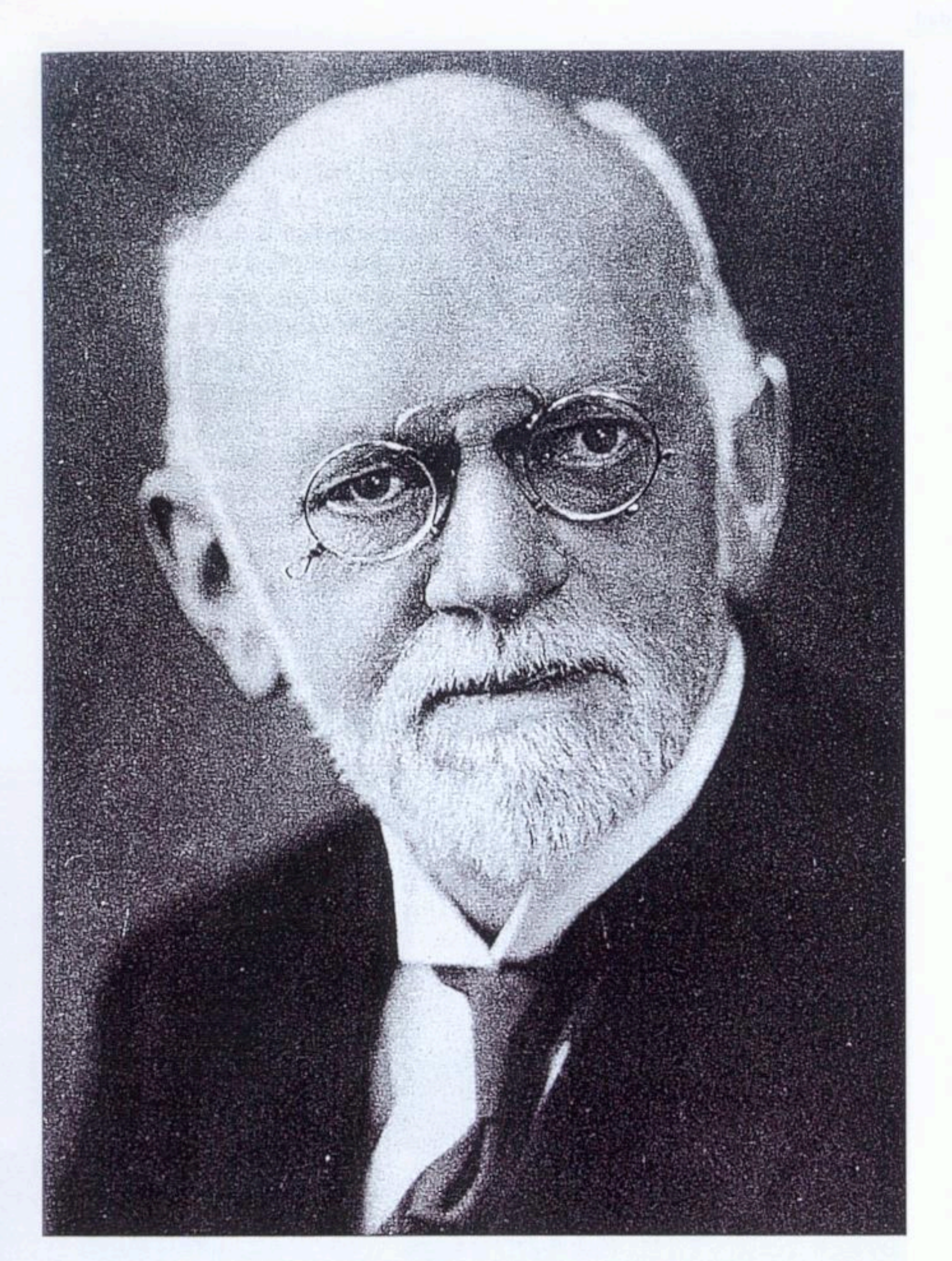
Nacido en 1862 cerca de Königsberg, capital de Prusia oriental, y fallecido en Göttingen, ciudad que, junto con otros, convirtió en el centro matemático más importante del mundo, David Hilbert levanta su enorme estatura en la frontera entre dos siglos y entre dos mundos. De una parte, la Alemania de Bismarck y al auge paralelo de la investigación y la industria; de otra, el régimen nazi y, a partir de 1933, la desertización de las universidades alemanas. Ya reconocido, e incluso famoso, a finales del siglo pasado, Hilbert ha sido, con Henri Poincaré, el matemático que una más amplia y profunda influencia ha ejercido sobre el desarrollo matemático de nuestro siglo. Fue precisamente en 1900 cuando, invitado a dar una de las conferencias generales del Congreso internacional de matemáticos celebrado en París, decidió «mirar hacia el futuro» y proponer un programa de investigación para el siglo. Presentó, en primer lugar, una reflexión sobre el papel de los problemas, sobre el rigor y la simplicidad de los métodos utilizados para resolverlos, sobre la necesidad de dar respuestas exactas (con un sí, un no o un imposible —como para la cuadratura del círculo o la prueba del axioma de las paralelas—); luego enunció veintitrés problemas abiertos cuya resolución encomendaba al «futuro». El «futuro» iba a trabajar en ellos durante medio siglo e incluso más.

Talentos múltiples. Hilbert no fue hombre de una sola especialidad. Del álgebra al análisis, pasando por la aritmética y la geometría, su ingenio se aplicó a todos los sectores de las matemáticas, a los que añadió los de la física, la lógica y la filosofía de las matemáticas. En todos ellos obtuvo resultados o escribió trabajos que aún hoy son piedras angulares u obras de referencia.

> Los «paseos matemáticos» propiciaron una exploración sistemática de la disciplina

Baste citar su famosa demostración de la existencia de una base finita para todo sistema de invariantes (1988), que cristalizó en célebres discusiones sobre la noción de existencia en matemáticas, su monumental informe sobre la teoría de números algebraicos (1897), en la que los aritméticos siguen inspirándose, su axiomatización de la geometría (1899), que ni siquiera un encarnizado defensor de la intuición podría ignorar, su teoría de los operadores lineales en un espacio con un número infinito numerable de dimensiones —nuestros «espacios de Hilbert»— (1900-1920), su tratado sobre los *Métodos matemáticos de la física* (1924, 1937), escrito con R. Courant, todavía útil a los estudiantes, su teoría de la demostración (1929-1930), que inauguró uno de los sectores principales de la lógica contemporánea, su epistemología finitista, cuyo «programa» sigue orientando el trabajo de los lógicos y alimentando la reflexión de los filósofos.

Königsberg, ciudad natal de Hilbert, pertenece a la historia de las matemáticas por sus puentes. El problema de los siete puentes de Königsberg, resuelto por Leonhard Euler en el siglo XVIII, fue el primer ejercicio de topología. Cuando Hilbert entró en la universidad, en otoño de 1880, K. Weierstrass y L. Kronecker eran los más grandes matemáticos alemanes; G. Cantor, profesor en Halle, estaba elaborando la teoría de conjuntos e introduciendo en matemáticas el infinito tal como actualmente lo conocemos. Weierstrass y Kronecker cimentaban el prestigio de Berlín, donde también estaban E. Kummer y H. von Helmholtz. Pero Hilbert no estudiaría en Berlín. Tras un semestre en Heidelberg con L. Fuchs, conocido por sus trabajos sobre ecuaciones diferenciales lineales, regresó a Königsberg para asistir a las lecciones de H. Weber sobre teoría de números, teoría de funciones y teoría de invariantes (polinomios homogéneos de varias variables que no cambian al someter dichas variables a cierta clase de transformaciones lineales). En Königsberg, Hilbert trabó gran amistad con H. Minkowski, cuyo talento matemático, pronto revelado, había sido ya distinguido con un premio de la Academia de ciencias de París (1883). Ambos frecuentaban a otro matemático precoz, A. Hurwitz, quien en 1884 sustituyó a Weber en la Universidad. Cada día, a las cinco de la tarde, Hilbert daba con sus dos amigos un paseo en el que se hablaba de matemáticas. Hilbert



Con Henri Poincaré, David Hilbert (foto adjunta) fue el matemático que mayor influencia ejerció sobre su disciplina durante la primera mitad del siglo XX. Así lo atestiguan los numerosos teoremas y resultados que llevan su nombre (espacio de Hilbert, teorema de los ceros de Hilbert, símbolo de Hilbert en teoría de números, etc.). También se hizo célebre al formular, en el congreso internacional de matemáticos de 1900, una lista de veintitrés grandes problemas que orientaron buena parte de las investigaciones futuras. (Foto fotoarchiv Städtisches Museum Göttingen)

gustaba mucho de esta forma animada y agradable de enseñar y aprender. Cuando fue profesor a su vez, convirtió esta costumbre en una memorable institución. Los «paseos matemáticos» le sirvieron siempre para explorar sistemáticamente con sus amigos y alumnos el campo de los problemas matemáticos. Partiendo de una sugerencia de C. von Lindemann, el matemático que al demostrar la trascendencia de π (la inexistencia de ningún polinomio de coeficientes enteros del que π fuera raíz) puso fin a los intentos de cuadrar el círculo, Hilbert acometió el estudio de la teoría de invariantes para su disertación de doctorado. En 1885 lo encontramos en Leipzig, donde conoció a F. Klein, autor del famoso programa de Erlangen. Este último unificó y clasificó las distintas geometrías, euclídea, no euclídea, proyectiva, etc., según el conjunto de transformaciones que dejan invariantes determinadas características; por ejemplo, en la geometría usual (la geometría euclídea), el grupo formado por las traslaciones, las rotaciones y las simetrías deja invariantes la longitud de los segmentos, la magnitud de los ángulos y la forma y el tamaño de las figuras del plano o del espacio. Klein, cuya fama había llegado por aquel entonces a América, reconoció en Hilbert al «futuro hombre de las matemáticas». Lo envió a París para que entrara en contacto con el genio de Poincaré (del que él se sentía celoso hasta el punto de deprimirse).

Hilbert rechazó el credo según el cual sólo los objetos construidos a partir de los enteros positivos tienen existencia legítima

Hilbert asistió al curso de teoría del potencial y de mecánica de fluidos de Poincaré, conoció a E. Picard y a C. Hermite, quien le indicó el problema irresuelto de P. Gordan sobre los invariantes cuya solución iba a hacer famoso a Hilbert. De regreso a Alemania, en junio de 1886, se detuvo en Göttingen, donde se había instalado Klein, y en Berlín, donde Kronecker lo recibió afectuosamente. Como Dedekind y Weber, Hilbert siguió los preceptos de Kronecker, aplicar los métodos aritméticos a los problemas planteados en otros sectores de las matemáticas, y envió regularmente sus primeros artículos a Kronecker. Hilbert admiraba a Kronecker y al tiempo desaprobaba su filosofía. Atento a los resultados de Cantor sobre los conjuntos infinitos, el «paraíso» del que no quería que los matemáticos fueran expulsados, no pudo aceptar el credo del maestro según el cual sólo tienen existencia legítima los objetos matemáticos que pueden construirse a partir de los números enteros positivos mediante un número finito de etapas. Tampoco pudo aceptar, por tanto, las objeciones de éste al uso por Weierstrass de los números irracionales representados por sucesiones infinitas de números racionales, ni sus ataques a la numeración transfinita de Cantor, que prolonga la sucesión de los números enteros con una jerarquía de nuevos números asociados a conjuntos infinitos.

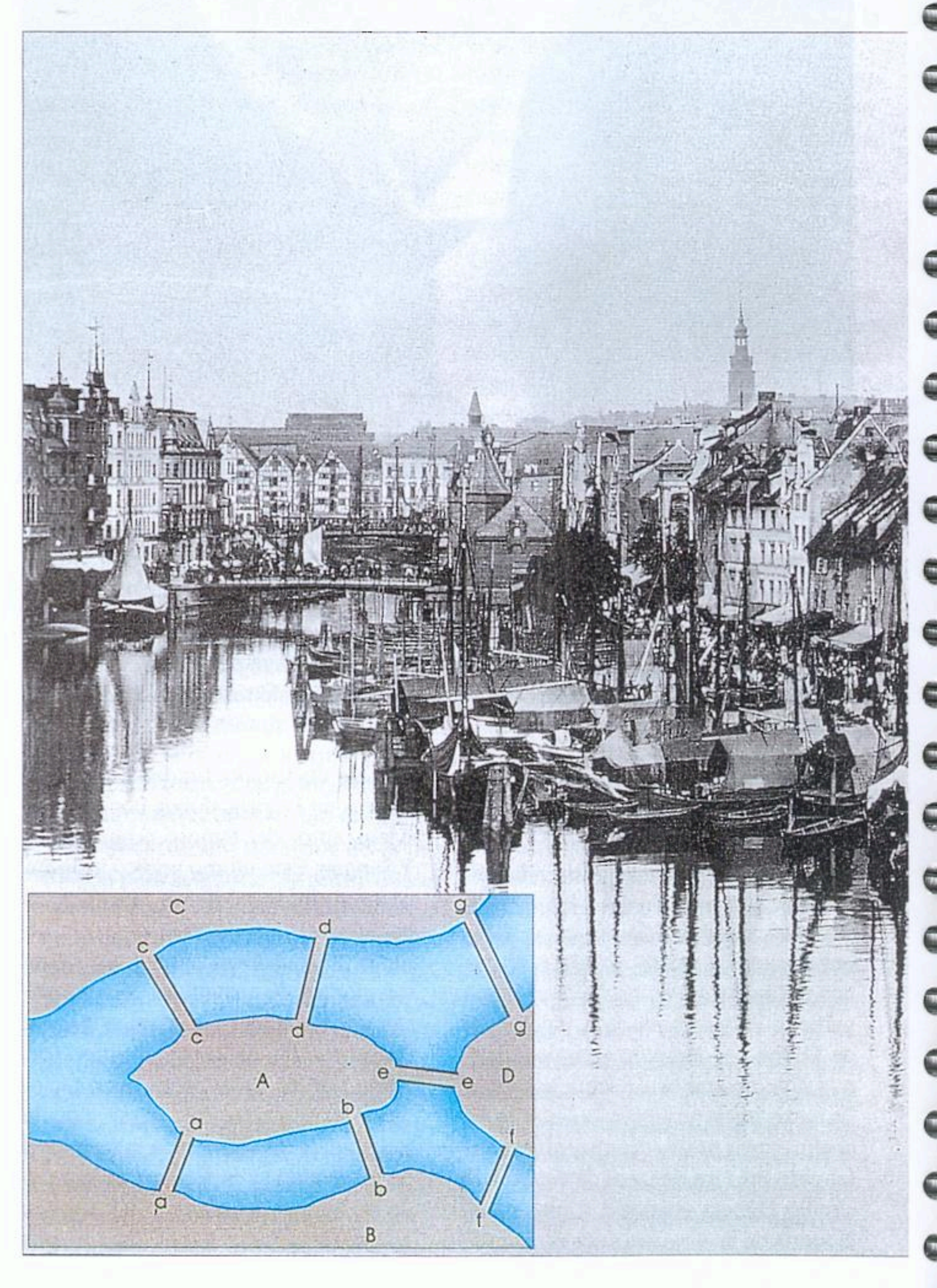
Unas pruebas desconcertantes. Tras su habilitación, Hilbert se fue a Erlangen a oír a Gordan y se dedicó al problema de éste: demostrar que todo sistema de invariantes tiene una base finita, es decir, que existe un número finito de invariantes, $I_1, I_2,..., I_n$ tal que todo invariante del sistema pueda es-

cribirse como función polinómica $P(I_1, I_2,..., I_n)$. Gordan sólo había conseguido calcular la base en el caso más simple (el de las «formas binarias» o polinomios homogéneos de dos variables). En 1988, Hilbert demostró de una forma general que sólo existe una base. No trató de construir efectivamente esta base ni de indicar un medio para construirla. Probó indirectamente su existencia demostrando que la hipótesis contraria llevaría a una contradicción. Así pues, para la existencia matemática bastaba una prueba de no contradicción. Éste fue el primer gran resultado de Hilbert. Algunos se sintieron desconcertados por esta prueba no constructiva; Lindemann la calificó de «extraña». Otros, como Kronecker y Gordan, se mostraron más agresivos; sabemos qué comentó este último: «*no son matemáticas, es teología*». Dicho de otro modo, hacía falta un acto de fe. Sólo Klein reconoció la genialidad del autor; la prueba le pareció «*perfectamente simple y lógicamente irresistible*». Más tarde, trató de convencer a Hilbert de que aceptara un nombramiento en Göttingen: «*Usted, por la orientación de su obra y la fuerza de su pensamiento, es el hombre que necesito... Cuento con usted para dar un nuevo impulso a nuestra Escuela matemática*». Klein, deseoso de organizar eficazmente la promoción de la ciencia, sabía lo que valía un talento superior.

Hilbert defendió contra el derecho vigente que prohibía a las mujeres acceder a los cargos de enseñanza superior

Indudablemente, Hilbert no fue el primer matemático que usó pruebas indirectas. Y tampoco desdeñó las pruebas constructivas, ya que en 1892 dio una nueva prueba, constructiva esta vez, de su teorema de la base finita. Pero para él las demostraciones generales de existencia, que permiten descubrir la estructura de los objetos considerados sin realizar cálculos, caracterizaban las matemáticas modernas. Por lo demás, Hilbert no sólo demostró el teorema de la base finita, sino también otros varios teoremas análogos, que son instrumentos esenciales para el estudio de los conjuntos algebraicos (definidos por sistemas finitos de ecuaciones polinómicas) y que iban a desempeñar un papel de primera magnitud en el desarrollo de la geometría algebraica. Gordan estaba en lo cierto al reconocer finalmente que «incluso la teología» (es decir, los métodos abstractos) tenía sus aspectos positivos.

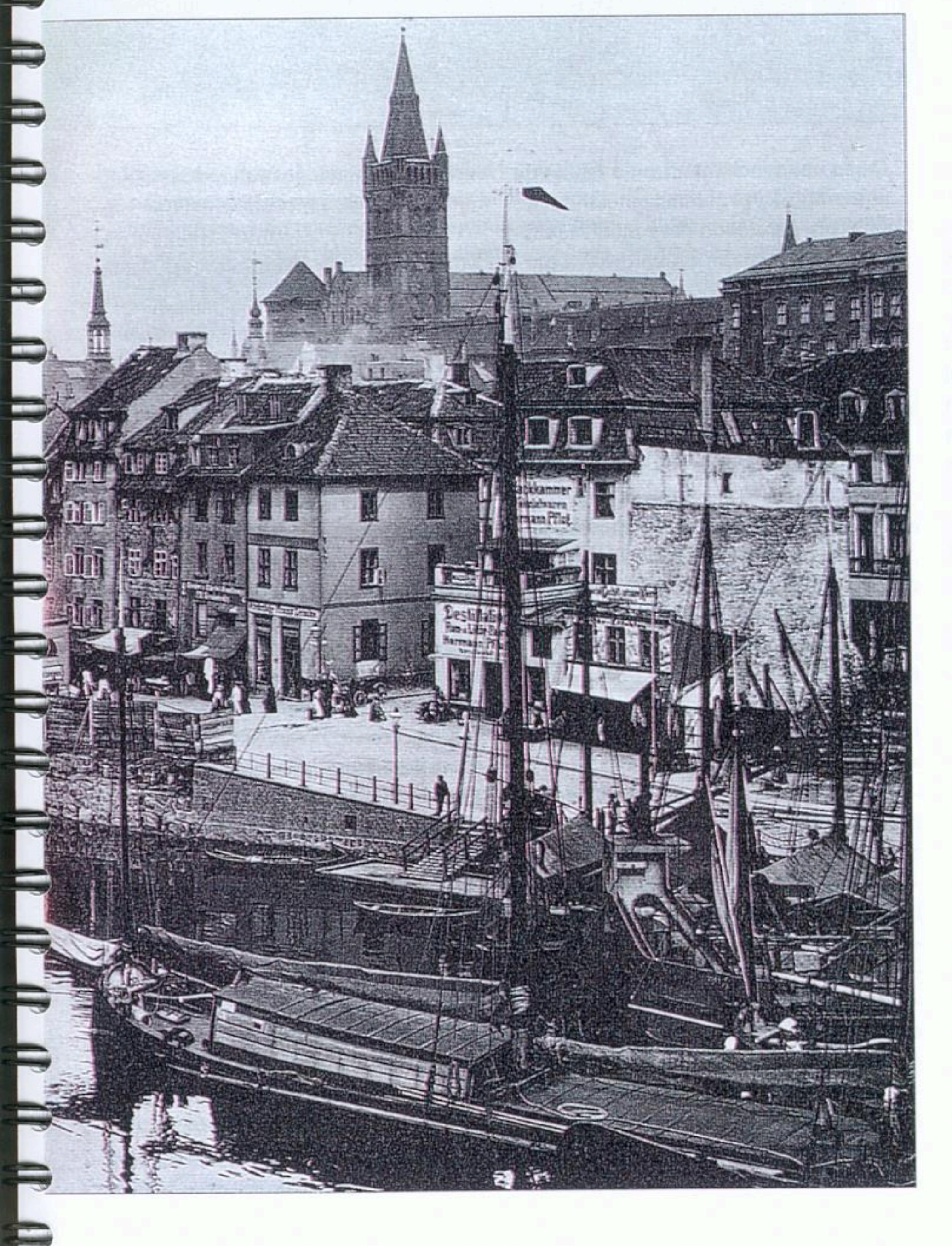
Los fastos de Göttingen. Hilbert llegó a Göttingen en marzo de 1895 para sustituir, a la edad de treinta y tres años, a H. Weber, destinado a Estrasburgo. La reputación de Göttingen era entonces centenaria, ilustrada por los nombres de Gauss, Wilhelm Weber, Dirichlet, Riemann y Dedekind. Desde su instalación en Göttingen (1886), Klein atrajo allí a numerosos estudiantes extranjeros, principalmente norteamericanos, y trató de estrechar lazos entre la ciencia y el mundo socioeconómico. Creó una asociación en la que físicos y matemáticos tenían la ocasión de conocer a jefes de empresa y, por iniciativa suya, aparecieron una serie de institutos tecnológicos que constituían otras tantas correas de transmisión entre la ciencia y la industria. Klein quería para Göttingen el liderazgo en Alemania y en todo el mundo; Hilbert se lo daría. En pocos años, se convirtió en el matemático alemán más conocido y en el igual de Poincaré. Göttingen hizo palidecer la estrella de Berlín, donde ni Fuchs, ni K.H.A. Schwartz ni G.F. Frobenius lograban contrarrestar el embrujo que ejercía sobre la juventud. Mentes valiosas sucumbie-



ron espontáneamente a este atractivo: H. Weyl, M. Born, E. Zermelo, B.L. van der Waerden, el autor del manual de álgebra moderna más utilizado desde 1930. Otros respondieron a una llamada explícita: Minkowski, C. Runge, E. Landau, Emmy Noether, a la que Hilbert defendió contra el derecho vigente que prohibía a las mujeres acceder a los cargos de enseñanza superior. Hilbert aprovechó repetidamente ofertas prestigiosas, como la cátedra de Sophus Lie en Leipzig o la de Fuchs en Berlín, para obtener, como premio de su negativa, la creación de puestos de enseñanza o de laboratorios nuevos en Göttingen.

Paso obligado. La Escuela de Hilbert irradió en el mundo entero. De todas partes llegaban alumnos, de Francia, donde más tarde se formaría el grupo Bourbaki, de Italia, de Países Bajos, de Grecia, de Rusia, de Japón, de Estados Unidos, etc. Hay que creer, como escribía Minkowski, que una estancia en Göttingen insuflaba el deseo de «hacer grandes cosas». La energía de Hilbert, su voluntad, su fe en el futuro y su confianza en la capacidad de la razón para hallar solución a todos los problemas suficientemente bien planteados, se transmitía a cuantos le rodeaban.

David Hilbert nació en 1862 cerca de Königsberg (actualmente Kaliningrado), capital en aquel entonces de la Prusia oriental. Esta ciudad es bien conocida por los matemáticos a causa del «problema de Königsberg» que el suizo Leonhard Euler estudió en 1736. Este último formulaba el problema en los siguientes términos: «En Königsberg, en Prusia Oriental, hay una isla A, llamada Kneihof, rodeada de un río que se divide en dos brazos [...] pero los brazos de dicho río son atravesados por siete puentes a,b,c,d,e,f,g. Se proponía el siguiente problema sobre los puentes: ¿puede alguien pasar una vez por cada puente, pero sólo una? Unos afirmaban que tal cosa era posible, otros lo negaban y otros aún lo dudaban, pero nadie lograba demostrarlo. Por mi parte, he transformado el problema en este otro, mucho más general: cualquiera que sea la figura del río y la distribución en brazos, y cualquiera que sea también el número de puentes, determinar si una persona puede atravesar el río pasando una sola vez por cada puente». Euler, entre otros autores, demostró que esto es imposible en el caso de los siete puentes de Königsberg. (Foto BPK)



En sus convicciones epistemológicas se aliaban el rigor y la simplicidad y se complementaban la lógica y la física

Se le iba a ver desde todos los horizontes de la ciencia. Con tan numerosos discípulos, su Escuela iba a desarrollar múltiples dominios de excelencia. En las manos de E. Noether, de E. Artin y de van der Waerden, la Escuela daría origen a nuestra «álgebra moderna».[1] E. Hecke y C. Siegel desarrollaron la tradición aritmética que inspiró los notables trabajos del francés André Weil. Born, con W. Pauli y W. Heisenberg, creó la escuela de mecánica cuántica. P. Bernays, W. Ackermann y G. Gentzen fueron, con Hilbert, los pilares de una escuela de lógica cuyos trabajos sirvieron de punto de partida al francés J. Herbrand, al austriaco Kurt Gödel y al polaco Alfred Tarski, y de alimento a la filosofía de los franceses J. Cavaillès y A. Lautman. Fuera físico o filósofo, matemático o lógico, ningún espíritu interesado en la ciencia de punta dejó de hacer, en el decenio heroico, su peregrinación a la ciudad de Gauss y de Riemann, convertida en faro de la modernidad y en «santo lugar del pensamiento puro». Para muchos, fue literalmente el descubrimiento de un «nuevo mundo» exaltante y magnético.[2]

Pero volvamos a los años 1895-1900.



Hilbert reanudó en Göttingen su costumbre de los paseos matemáticos, a los que llevaba a sus alumnos más brillantes, a sus «niños prodigio», como les llamaba. Dictó cursos sobre los temas más variados, feliz de seguir desbrozando nuevos campos y deseoso de mostrarse digno de las esperanzas que en él había depositado Klein. Sus lecciones, cuidadosamente preparadas, nada tenían de rígido. Hilbert no recitaba bloques enteros de saber, más bien daba la impresión de buscar y encontrar a medida que iba exponiendo. El auditorio creía asistir a veces al proceso mismo de la creación matemática, lo que indica que la concentración y el fervor de Hilbert debían de ser considerables.

Números algebraicos. En sus primeros años en Göttingen, Hilbert se concentró en su *Zahlbericht*, un informe encargado por la sociedad matemática alemana que versaba sobre la teoría de números algebraicos (las raíces de un polinomio de coeficientes enteros). La obra, una auténtica joya, era una exposición sistemática y unificada de los resultados obtenidos desde Kummer, con codificación de los términos y simplificación de las pruebas, que sometió al agudo escrutinio de su amigo Minkowski. Hilbert quería preparar el futuro y facilitar el descubrimiento. Lo consiguió con creces: *el zahlbericht* fundaba la teoría de las extensiones abelianas de un cuerpo de números algebraicos (extensiones, que satisfacen ciertas propiedades, de una extensión finita del cuerpo de números racionales), llamada «teoría del cuerpo de clases», y formulaba conjeturas que el japonés T. Takagi y el francés C. Chevalley, ambos discípulos de Artin, demostrarían más tarde; por otra parte, su teorema nº 90 es una raíz de nuestra álgebra homológica (estudio de los invariantes para objetos de naturaleza algebraica o algebraico-topológica). Según los especialistas, el *Zahlbericht* sigue siendo una mina.

Después de la teoría de invariantes y de números algebraicos, Hilbert se orientó hacia el análisis y abordó el problema de Dirichlet. Se trata de encontrar, para un dominio «abierto y acotado» D del plano, una función *u*, continua en D (es decir, que en todo punto de D se cumpla $\partial^2 u/x\partial^2 + \partial^2 u/\partial y^2 = 0$) y cuya restricción a la frontera de D sea una función continua f dada de antemano. El problema, inicialmente, se había planteado para medir la distribución del calor en un disco a partir de los valores conocidos en su frontera, pero otros fenómenos, eléctricos o hidrodinámicos, pueden recibir el mismo tratamiento matemático. La existencia de u se daba por cierta sobre la base de consideraciones físicas. P.G. Dirichlet (en un texto póstumo publicado en 1876) había conjeturado que, de entre todas las funciones continuas que toman los valores definidos por f en la frontera de D, u era aquella para la cual la integral siguiente, calculada en dominio D, alcanzaba su mínimo:

$$\iint \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right) \partial x \, \partial y$$

Hilbert reformuló el problema en el marco del cálculo de variaciones y demostró en 1899 la conjetura, llamada por Riemann «principio de Dirichlet». Riemann lo utilizó en su disertación de 1851, Principios fundamentales para una teoría general de funciones de una variable compleja, y Weierstrass le opuso un contraejemplo, de tal modo que los matemáticos se veían obligados a hacer contorsiones para llegar a los resultados de Riemann sin pasar por el principio de Dirichlet. Mediante ciertas restricciones sobre la frontera de D y la función *f*, Hilbert levantó las objeciones de Weierstrass y legitimó totalmente la disertación de Riemann, una de las obras más importantes de la historia de las matemáticas modernas. Ilustraba así una de sus más profundas convicciones epistemológicas: el rigor (exigencia de prueba) y la simplicidad son aliados, no enemigos. O, si se prefiere, la lógica y la física son complementarias y no opuestas. Klein estaba lleno de admiración y M. von Laue, futuro premio Nobel de física (en 1914), quedó muy impresionado por el curso de cálculo de variaciones que Hilbert dictó aquel año.

Espacios hilbertianos. Hilbert se orientó pronto hacia el análisis de las ecuaciones integrales (ecuaciones que contienen términos en los que la función desconocida aparece bajo el signo de integración) y trató de realizar el programa de Poincaré de una teoría que unificara distintos aspectos de la física y del análisis matemático. Deseoso como siempre de trabajar para las generaciones futuras, Hilbert, como en teoría de números algebraicos, reexpuso y simplificó con métodos nuevos los resultados conocidos.

En esta ocasión, Hilbert desarrolló sistemáticamente la teoría de las formas cuadráticas (o polinomios homo-



El primer gran éxito de D. Hilbert tuvo lugar en 1888, cuando logró resolver un problema planteado por el matemático P.Gordan: demostrar que «todo sistema de invariantes posee una base finita». La demostración de Hilbert era no constructiva. No permitía construir la base, se limitaba a probar su existencia sentando la hipótesis contraria y demostrando que llevaría a contradicción. Por su carácter no constructivo, la demostración de Hilbert desasosegaba a muchos de sus colegas. Pero F. Klein, que llamaría a Hilbert a Göttingen en 1895, supo reconocer su genialidad. Vemos aquí a Hilbert (segundo de arriba por la izquierda) durante una reunión de matemáticos alemanes celebrada en Bremen en 1890. Se advierte la presencia entre ellos de celebridades como H. Minkowski, H. Weber, G. Cantor, P. Gordan y F. Klein.
(Foto Mathematisches Forschungsinstitut Oberwolfach)

géneos de segundo grado) de infinitas variables: fue el punto de partida de los espacios de Hilbert y de la teoría espectral, así llamada por el propio Hilbert. Este definió el espacio de sucesiones de cuadrado sumable (sucesiones en las que la suma de los cuadrados de los términos es convergente) y resolvió los problemas planteados interpretando las ecuaciones en términos de transformaciones lineales de dicho espacio. A partir de ahí, un lenguaje geométrico se impondría a los analistas. J. von Neumann y F. Reisz axiomatizarían la teoría de los espacios de Hilbert, que pasó a ser una importante herramienta de la física matemática. En 1912, Hilbert reunió todos sus trabajos sobre las ecuaciones integrales en un libro que sentaba las bases del análisis funcional (estudio de espacio de funciones) actual.

Los números irracionales existen. En el cambio de siglo, el genio de Hilbert, indiscutido, era realmente diverso. En 1899, el año de la publicación del principio de Dirichlet, Hilbert publicó el libro que al hacer —una vez más— la síntesis de los trabajos anteriores, consagró de manera definitiva la aparición del método axiomático. En *Los fundamentos de la geometría*, Hilbert clasificó los distintos axiomas de la geometría y fue uno de los primeros en considerar el espacio como un concepto matemático y sólo matemático, y no como el lugar, la forma o la estructura de nuestra experiencia. La geometría se convirtió en una ciencia pura, como el álgebra o la aritmética. Ya no era la expresión idealizada de la realidad sensible como pensaba Aristóteles (*Metafísica*, libros M y N) ni la construcción de las leyes formales de nuestra percepción del mundo, como sostenía Kant (*Crítica de la razón pura*, Estética trascendental) ni, como aseguraba Riemann (*sobre las hipótesis que sirven de base a la geometría*, publicado póstumamente en 1867), el conjunto de hipótesis previas a nuestro conocimiento de la realidad, sobre cuyo estatus epistemológico no se dice nada, ni siquiera los hechos que sirven de base a la geometría, como pensaba Helmholtz (1868). El vínculo con la realidad sensible quedaba simplemente roto. Se hablaba de puntos, de líneas, o de planos como se podía haber hablado de mesas, de sillas o de jarras de cerveza. No se trataba de saber qué proposiciones parecen verificadas en el mundo que nos rodea, sino cuáles son indispensables para la demostración de tal o cual proposición geométrica dada. Se hacía hincapié en las relaciones lógicas de compatibilidad y de dependencia entre proposiciones: dos proposiciones son lógicamente compatibles si se verifican simultáneamente en un mismo *modelo*, es decir, en un dominio que consiste en un conjunto de elementos que verifican ciertas relaciones dadas desde un principio. Una proposición M es independiente de otras si existe un modelo que verifica éstas y la negación de M: así se puede demostrar la independencia del axioma euclídeo de las paralelas que generaciones de matemáticos habían tratado en vano de deducir de otros axiomas. Inmediatamente valorado como un clásico, *Los fundamentos de la geometría* de Hilbert conoció diez ediciones en su lengua original y múltiples traducciones. Es una magnífica muestra del espíritu de las matemáticas modernas, axiomatizadas y abstractas.

La geometría se convirtió en una ciencia pura, como el álgebra o la aritmética

El interés de Hilbert por la lógica derivaba de su antigua percepción de la no contradicción como resorte de las pruebas de existencia y de su axiomatización reciente de la geometría. Se añadía a ello la voluntad de arruinar la prohibición de Kronecker sobre los conceptos y métodos que recurren al infinito y de poner coto al efecto devastador de las paradojas que el alemán *E.* Zermelo y el británico B. Russell descubrieron hacia 1904 en la

teoría de conjuntos. Es conocido el ejemplo famoso del conjunto E de los conjuntos que no pertenecen a sí mismos: se comprueba fácilmente que $E \in E$ si y sólo si $E \notin E$.

A Lindemann, que le hablaba de su prueba de la trascendencia de π, Kronecker le había respondido lo siguiente: «*¿De qué sirve todo esto? Los números irracionales no existen*». Hilbert, por su parte, quería demostrar que estos números existían. Y probarlos de acuerdo con las exigencias de Kronecker, a las que en aquel momento consideró como «una necesidad filosófica universal de nuestro entendimiento»: en un número finito de pasos y a partir de un número finito de hipótesis. Hilbert acabó, pues, adoptando la posición de Kronecker,

Lo más sorprendente es que el desarrollo masivo de las matemáticas se hizo sin cambio radical de las problemáticas

al menos en lo referente a las necesidades de la prueba. Pero persistió en su opinión de que una entidad existe tan pronto como se ha demostrado que no implica ninguna contradicción. Trató, por lo tanto, de demostrar que el sistema (finito) de axiomas que definen los números reales es no contradictorio. Mejor aún, ya que los reales pueden ser considerados como límites de sucesiones de números racionales y que éstos son equivalentes a pares de enteros (en efecto, pueden escribirse en la forma p/q, donde p y q son enteros y q es no nulo), bastaba demostrar la no contradicción de los axiomas de la aritmética de los números enteros. Este fue el contenido del segundo problema enunciado por Hilbert en 1900. Minkowski valoró la importancia de la propuesta: plantear como problema algo de lo que los matemáticos nunca habían dudado.

Los veintitrés problemas. Hilbert acometió entonces una remodelación simultánea de los fundamentos de la aritmética y de la lógica que dio lugar a una nueva disciplina, la «metamatemática» o teoría de la demostración, y generó una actitud epistemológica general, el punto de vista «finitista». Este pretendía asegurar con medios finitos el derecho de operar con el infinito. Los «medios finitos» consisten en el uso combinado de la axiomática de la teoría de la demostración.[3] Las demostraciones son sucesiones finitas de fórmulas; Hilbert pensaba que así el estudio podía mantenerse lo más cerca posible de la intuición y que de esta manera una metamatemática finitista podía garantizar las matemáticas del infinito. Se trataba en cierto modo de poner el espíritu de Kronecker al servicio de las entidades y los conceptos de Cantor.

Gödel demostraría en 1931 la limitación inherente a la supeditación de lo infinito a lo finito; incluso para la aritmética elemental, y antes siquiera de llegar a lo transfinito propiamente dicho, ningún número finito de axiomas reduce a un sí o a un no la respuesta a toda cuestión aritmética. Así, el conjunto de las proposiciones verdaderas para los números enteros no es reducible, vía inferencia lógica, a un conjunto finito de axiomas. Ello convierte en caducas las cuestiones de no contradicción sin por ello liquidar el espíritu finitista. Muy al contrario, las empresas de constructivización de las matemáticas clásicas empezaron a florecer. Algunas, todavía hoy, apelan insistentemente al finitismo de Hilbert. Así, las «*reverse mathematics*» de los norteamericanos H. Friedman y S.G. Simpson, que plantean la cuestión de saber qué axiomas de existencia conjuntistas son necesarios para demostrar los teoremas de las matemáticas ordinarias, donde por matemáticas ordinarias se entiende la teoría de números, la geometría, el cálculo diferencial, el análisis real y complejo, la combinatoria, el álgebra numerable los espacios de Banach separables, la teoría de la calculabilidad y la topología de los espacios métricos completos y separables, y, por matemáticas no ordinarias, el análisis funcional abstracto, la teoría abstracta de conjuntos, el álgebra universal y la topología general. El resultado esencial, que data de 1977-1979, es la construcción de un sistema lógicamente bastante débil aunque matemáticamente lo bastante potente como para expresar importantes teoremas clásicos, notable ilustración actual de la epistemología hilbertiana.

Hilbert debe su fama no sólo a los numerosos resultados que obtuvo, de los que hemos tratado de dar una idea general, sino también a los veintitrés problemas que propuso en el congre-



El instituto de matemáticas de la Universidad de Göttingen acogió a David Hilbert en 1895. La ciudad gozaba ya de una sólida reputación científica, puesto que grandes científicos como C.F. Gauss y B. Riemann habían trabajado en ella. Primero bajo el impulso de F. Klein, instalado allí desde 1886, y luego gracias a Hilbert, Göttingen se convirtió en pocos años en el centro mundial de las matemáticas. La irradiación de la escuela de Hilbert atrajo a talentos procedentes del mundo entero. Pero a raíz de las persecuciones emprendidas por los nazis, llegados al poder en 1933, la Escuela de Göttingen perdió rápidamente a sus miembros y su prestigio decayó, con gran pena de Hilbert. (Foto Universitatarchiv Göttingen)

De la Conferencia de 1900 a las del año 2000

La Unión matemática internacional ha tomado una iniciativa que pretende adaptar dicha conferencia al desarrollo actual de las matemáticas. El anuncio se hizo público en una reunión de su Comité ejecutivo en Río de Janeiro (mayo de 1992). Por iniciativa del presidente de la Unión, el francés Jacques-Louis Lions, el año 2000 ha sido reconocido por las autoridades internacionales como «Año mundial de las matemáticas».

Se han puesto en marcha programas en distintas direcciones: enseñanza, países en vías de desarrollo, investigación. El análogo estricto de la conferencia de 1900 será probablemente una serie de conferencias dadas en varios lugares del mundo, una idea propuesta por la delegación norteamericana en el congreso de Kyoto de 1990. Los matemáticos, por tanto, están dispuestos a aceptar el reto de una prospectiva de su disciplina teniendo en cuenta que la actividad matemática se ha globalizado. No obstante, con el siglo XXI, los matemáticos están en una situación muy distinta que a principios de siglo a causa del importante papel social que su disciplina ha pasado a desempeñar. Un antídoto al cientificismo, que puede volver a tentar a algunos matemáticos, podría consistir en confrontarlos a dicho papel, por ejemplo, mediante un análisis cuidadoso del uso de las matemáticas en economía o en biología. Pero ello no significa que su desarrollo deba estar determinado por este papel social. Aunque difícil de hacer admitir, esta exigencia de una autonomía del pensamiento matemático es indispensable para dejar espacio a la creación. Ahora más que nunca, es imposible reducir las perspectivas de desarrollo de las matemáticas a la resolución de una lista de problemas, aunque la propongan matemáticos de excepción.

so de 1900 y que ejercieron una gran influencia sobre los matemáticos del siglo XX.[4] ¿En qué medida estos problemas siguen teniendo actualidad? No es posible precisarlo sin establecer una tipología de dichos problemas y presentar sucintamente su orden de aparición en la conferencia. Hay que destacar que constituyen una elección, basada en la ejemplaridad de los métodos, y en modo alguno representan el conjunto de los campos de interés de Hilbert. En 1900, por ejemplo, Hilbert estaba en pleno proceso de maduración del análisis funcional, pero de este tema no se habla en los veintitrés problemas.

Casi todos tienen un enunciado preciso; la única excepción es el vigesimotercero y último, que versa sobre cálculo de variaciones. En la mayoría de los casos, se trata de problemas situados en la confluencia de subdisciplinas matemáticas que importa abordar de varias maneras. Esta importancia dada a las interficies «internas» de las matemáticas sigue vigente; el desarrollo de las matemáticas en el último decenio procede sobre todo de tales cruces.

No todos los problemas planteados por Hilbert tiene la misma importancia. Algunos pueden ser calificados de «grandes problemas», otros son más particulares. El décimo, por ejemplo, que trata de la resolución de ecuaciones de números enteros, contiene todas las cuestiones matemáticas cuya formulación puede reducirse a una ecuación algebraica; así lo demostró en 1970 el soviético I. Matijasevic. Al conjunto formado por los dos primeros problemas, que versan sobre lógica y fundamentos de las matemáticas (la no contradicción de los axiomas de la aritmética) se lo suele conocer como «programa de Hilbert». Como ya hemos dicho, el teorema de Gödel (1931) demuestra que no se puede abordar el infinito en matemáticas al modo restrictivo de Hilbert. Pero este estado de cosas no es lo bastante conocido (como se puede comprobar fácilmente leyendo ciertos programas de inteligencia artificial, que se olvidan de tomar en consideración las restricciones matemáticas derivadas de elecciones previas de reglas de prioridad). Notemos, sin embargo, que la consideración del infinito que propone Hilbert es precisamente la que está implementada en un ordenador.

Los cuatro últimos problemas tratan del cálculo de variaciones. Este campo de las matemáticas trata de definir una curva, una superficie o un objeto matemático más general que verifique una cierta propiedad de extremalidad (por ejemplo, un camino más corto, una superficie de área mínima, un forma dotada de resistencia mínima, etc.). En el siglo XX, este tema ha tenido una vitalidad extraordinaria, por lo que la cuestión planteada por Hilbert no guarda proporción con los desarrollos actuales. No obstante, Hilbert había tenido el «olfato» de terminar su texto con una cuestión general acerca de este campo. No sólo se han obtenido unos resultados extremadamente numerosos; también, y sobre todo, el punto de vista variacional ha invadido muchos dominios, como la geometría y el análisis, y está desempeñando un papel determinante en las aplicaciones industriales de las matemáticas.

Hacia 3000 a.C., los signos numéricos se organizan en una docena de sistemas diferentes. Uno designa cantidades discretas, otro unidades de superficie...

Una física fecunda. En su sexto problema, Hilbert se interesó por el tratamiento matemático de los axiomas de la física. Para él, esta cuestión formaba parte integrante de la actividad matemática. Dicho punto de vista, que no ha dado lugar a excesivas aplicaciones porque no ha prevalecido mucho en este siglo, regresó con fuerza en los años 1980, pues ciertas ideas tomadas de la física se mostraron extremadamente fecundas para resolver problemas puramente matemáticos (utilización de los instantones o de los monopolos para sugerir nuevos invariantes de variedades diferenciales, teoría cuántica de campos para calcular ciertos invariantes de nudos por medio de las integrales de Feynman, etc.). Este vaivén creativo entre las matemáticas y otras ciencias sigue siendo un problema epistemológico al orden de día, aunque haya sido y siga siendo objeto de numerosas reflexiones. Es indudable que ciertas cuestiones filosóficas planteadas a los matemáticos y a los físicos teóricos son particularmente profundas y requieren, para su reso-



David Hilbert clasificó los distintos axiomas de la geometría.

lución, la creación de conceptos radicalmente nuevos.

Los sectores de la ciencia cuyas relaciones con las matemáticas se están desarrollando actualmente son mucho más numerosos que en tiempos de Hilbert por la sencilla razón de que ciencias como la biología o la informática distaban, en aquel entonces, de haber alcanzado su nivel actual. No hay que olvidar, sin embargo, que ya a mediados del siglo XX Helmholtz se interesaba por la modelización de la visión, de la formación de las nubes y del oído (recuérdese también que Riemann escribió una «Mecánica del oído»). Estas perspectivas no estaban presentes en la problemática de Hilbert. La idea de la modelización, en el sentido moderno de una teoría matemática creada para dar cuenta de un conjunto de fenómenos, no figura en el texto de la conferencia de 1900.

Acerca de las relaciones entre las matemáticas y las demás ciencias, Hilbert y Poincaré, que a menudo se consideran opuestos, tenían puntos de vista bastante convergentes: «*Es hacia la naturaleza hacia donde hay que dirigir el grueso de nuestro ejército*». El progreso de las matemáticas, ¿viene de motivaciones internas o de solicitaciones externas? La cuestión sigue estando de candente actualidad y la respuesta que se impone parece ser «más o menos de ambas», por parodiar a Poincaré, para quien había sólo «*problemas más o menos resueltos*». Sin pretender ser iconoclastas, tenemos que decir que ya no es posible caracterizar el siglo XX como el siglo de Hilbert y Poincaré. Este siglo, en efecto, ha conocido un desarrollo masivo de las matemáticas. A título de ejemplo, había 226 matemáticos en el congreso ante el cual dio Hilbert su conferencia en 1900; en 1990, en Kyoto, había 6 000 asistiendo al vigesimoprimer congreso internacional de matemáticas. Este cambio de escala supone modificaciones dramáticas de la organización interna de la comunidad matemática y también de su misión, con un papel incrementado de la economía, la industria y los servicios. Por otra parte, las matemáticas se llevan la mejor parte en la formación. Pese a una gran diversidad de concepciones en los distintos sistemas educativos, todos los países han decidido otorgar a las matemáticas un gran peso en la enseñanza general; correlativamente, se ha producido una gran necesidad de enseñantes de esta disciplina. Por supuesto, la evolución de nuestro siglo hacia una sociedad de la información han contado mucho en el reconocimiento y la persistencia de esta necesidad.

Hilbert creía en el progreso y en el triunfo, por encima de todo, de la verdad científica

Matemáticas sin fronteras. Pero lo que más sorprende es que este desarrollo masivo se ha hecho sin cambio radical de las problemáticas. Se puede atribuir esto a una extraordinaria resistencia del fondo histórico de las matemáticas a solicitaciones exteriores. Lo que lo ha hecho posible, indiscutiblemente, ha sido la corrección de los conceptos que nos han legado generaciones anteriores de matemáticos de la talla de Hilbert. El chino Shiing Shen Chern, una figura histórica de las matemáticas del presente siglo, prefiere ver ahí una consecuencia de que «los conceptos fundamentales son raros».

Hilbert fue visto por sus contemporáneos como un espíritu dotado de una vitalidad y una capacidad lógica sin precedentes, desdeñoso de tradiciones y convenciones y defensor audaz de la libertad del pensamiento matemático. Heredero del siglo de las Luces, del que Königsberg, la ciudad natal de Hilbert, llevaba la huella dejada por la obra monumental del filósofo Immanuel Kant, Hilbert creía en el progreso y en el triunfo, por encima de todo, de la verdad científica. En el congreso internacional de matemáticas que siguió a la primera guerra mundial (Bolonia, 1928), Hilbert insistió en el carácter universal de las matemáticas: «*Todas las fronteras, sobre todo las nacionales, son contrarias a su naturaleza.*» La llegada de Hitler al poder y la persecución de los judíos lo conmocionó; asistió impotente a la «sangría», a la obligada partida de Courant, Emmy Noether, Born, Weyl y tantos otros.[5] Al nuevo ministro de Educación que a fines de 1933 le preguntaba por la salud de las matemáticas en Göttingen, «*ahora que han sido limpiadas de la influencia judía*», le respondió Hilbert: «*¿Matemáticas en Göttingen? ¡Pero si ya no hay!*».

HOURYA SINACEUR es directora de investigación del CERN, del que preside la 35ª sección (Pensamiento filosófico —Ciencia de los textos— Creación artística, científica y técnica). Trabaja en el Instituto de historia y filosofía de la ciencia y de la técnica de París.

JEAN-PIERRE BOURGUIGNON, director de investigación del CNRS, enseña actualmente en la Escuela politécnica de Palaiseau. Realiza su trabajo de investigación en el Centro de matemáticas de esta Escuela (URA 169 del CNRS, unidad de la que es director).

(1) Sobre el nacimiento del álgebra moderna, véase la segunda parte del libro de H. Sinaceur, *Corps et modèles. Essai sur l'histoire de l'algèbre réelle*, Vrin, 1991.

(2) D. Rowe, «Klein, Hilbert and the Göttingen Mathematical Tradition», *Osiris*, 2nd series, 5, 186, 1989.

(3) H. Sinaceur, «Du formalisme à la constructivité: le finitisme», *in Revue internationale de philosophie*, otoño de 1993.

(4) F.E. Browder (ed.), «Mathematical Developments arising from Hilbert Problems», *Proceedings of Symposia in pure Mathematics*, vol. XXVIII, American mathematical Society, 1976.

(5) N. Schappacher, «Questions politiques dans la vie des mathématiques en Allemagne (118-1935)», in *La Science sous le Troisième Reich* (sous la direction de Josiane Olff-Nathan), Seuil, 1993, p. 51.

PARA MÁS INFORMACIÓN:

«Mathematische Problema», en Hilbert, Gesammelte Abhandlungen, vol. III, Springer, 1935; reprint New York, Chelsea, 1965; trad. francesa en E. Duporcq, Compte rendu du deuxième congrès international des mathématiciens, Gauthier-Villars (1902).

Artículo «Hilbert» de la Encyclopaedia Universalis, 1985.

F. Corbalán, *La matemática en sus personajes*, Ed. Nivola, Madrid, 2000.

Los números *p*-ádicos

Daniel Barsky y Gilles Christol

A principios de siglo, el matemático Kurt Hensel inventó los números *p*-ádicos. ¿Qué designa este curioso vocablo? Designa unos números abstractos y difíciles de representar, pero también unas entidades que permiten a los especialistas en teoría de números construir unos potentes instrumentos de estudio.

Cuánto mide la diagonal de un cuadrado de un metro de lado? Al unísono, los estudiantes de secundaria responden sin vacilar ¡$\sqrt{2}$ metros! Y dan esta respuesta sin pensar que $\sqrt{2}$, y los números que se le parecen, han sido considerablemente molestos para los matemáticos, desde la Antigüedad hasta una época bastante reciente. Los números enteros y las fracciones que éstos permiten formar fueron fácilmente aceptados, pero no cabe decir lo mismo de los llamados números irracionales como $\sqrt{2}$ o π, cuyo desarrollo decimal es ilimitado y no periódico (por ejemplo, $\pi = 3{,}14159$). Hubo que esperar al siglo pasado para que los números irracionales fueran definidos de un modo totalmente satisfactorio con la construcción del conjunto de los números *reales* R a partir del cuerpo Q de los números *racionales* (que comprende los enteros y los cocientes de enteros). El conjunto Q de los números racionales era, por varias razones, incompleto. Ya hemos aludido a una de ellas: como muestra el ejemplo de la diagonal del cuadrado, no hay suficientes números racionales para representar todos los puntos de una recta. Para dar un fundamento sólido a todo el análisis matemático (límites, funciones, integración, ecuaciones diferenciales, etc.) y hacerlo progresar, era indispensable añadir los números irracionales a los números racionales para formar el cuerpo R de los números reales (véase recuadro «Cuerpo, distancia, valor absoluto»).

Un cuerpo menos intuitivo. En 1902, sin embargo, el matemático alemán Kurt Hensel inventó unos objetos, los «números *p*-ádicos», que constituyen una forma distinta de completar el conjunto de los racionales. El cuerpo de los números p-ádicos no se parece al de los números reales y es mucho menos intuitivo. Entre otras cosas, los números *p*-ádicos no se prestan a una interpretación geométrica tan simple como los números reales o los números complejos (fig. 1). Entidades relativamente abstractas, han tardado en encontrar su utilidad: hoy en día, sin embargo, los números p-ádicos ocupan un lugar central en muchas ramas de las matemáticas, como la teoría algebraica de números (estudio de las raíces de polinomios de coeficientes enteros) o la geometría algebraica (estudio de las soluciones de las ecuaciones polinómicas de varias variables).

Los números *p*-ádicos fueron introducidos para representar los números algebraicos en forma de series de potencias de un número primo *p*

¿Qué son pues los números *p*-ádicos? Hay varias maneras de definirlos. El método original, debido a Hensel, recurría a los números *algebraicos*, soluciones de ecuaciones polinómicas de coeficientes enteros. Hensel introdujo los números *p*-ádicos al tratar de representar los números algebraicos en forma de series de potencias de un número primo *p*. Pero el camino histórico, aunque permite comprender mejor el contexto en el que fueron inventados los números *p*-ádicos, no es el más fácil de presentar. Hemos elegido otro, más parecido a la construcción del cuerpo R de los números reales. Pero antes son necesarios algunos preámbulos. En particular, es necesario precisar las nociones de «valor absoluto» y de «distancia».

Varias distancias. Para un número racional x, el «valor absoluto» usual de x se representa por $|x|$ y vale simplemente x si el número es positivo y $-x$ si es negativo; en otros términos, el valor absoluto de x es siempre positivo: $|3| = 3$, $|-12| = 12$, $|-4/5| = 4/5$, etc. El valor absoluto de x puede interpretarse como la distancia entre x y 0. En general, la distancia entre dos números racionales x e y es el número $|x - y|$. Por ejemplo, la distancia entre -4 y $+8$ vale 12. Hay que hacer una observación de gran importancia para lo que sigue: el valor absoluto y la distancia están dotados de unas propiedades que cabe formular en términos muy abstractos (véase recuadro «Cuerpo, distancia, valor absoluto»). Por ello, en un determinado conjunto suele ser posible definir una o varias «distancias» que no necesariamente se parecen a la distancia geométrica e intuitiva a la que estamos acostumbrados. Lo veremos con la llamada «distancia *p*-ádica», pero antes deberemos explicar de qué modo la noción de distancia permite «completar» el cuerpo de los números racionales.

El valor absoluto y la distancia definidos más arriba son imprescindibles para el análisis. Por ejemplo, estas nociones permiten dar un sentido preciso a la afirmación «el número 5,12 está próximo al número 5,11» o al enunciado «la sucesión $u_n = 1/n$ tiende a cero cuando n tiende a infinito». Consideremos ahora la sucesión $(1 + 1/1)^1, (1 + 1/2)^2, (1 + 1/3)^3, (1 + 1/4)^4$, etc. Estos números racionales valen respectivamente 2, 2,25, 2,37,..., 2,44..., etc. y se van acercando (¡en el sentido de la distancia definida antes!) a un cierto límite e. Se puede demostrar que este límite no es racional: e no puede escribirse bajo la forma a/b, con a y b enteros.

Unas sucesiones que se aproximan. Por múltiples razones (comodidad, rigor, coherencia, etc.), los matemáticos deseaban completar el cuerpo Q de los números racionales incorporando números como e, es decir, límites de sucesiones de números racionales. El procedimiento es técnico pero bastante simple. Baste decir que consiste en razonar sobre sucesiones de números racionales que se van acercando unos a otros (aquí es donde interviene la distancia) y que, en cierto modo, identifica los números reales con todas las sucesiones posibles que gozan de esta propiedad. El conjunto **R** así construido contiene, además de los números racionales, los llamados números irracionales, aquellos cuyo desarrollo decimal consta de una infinidad de cifras detrás de la coma sin repetición periódica (por ejemplo $\pi = 3,14159265358...$ o $e = 2,718281828459...$).

Para encontrar el valor absoluto *p*-ádico de un entero positivo *n* basta dividir *n* por *p* y sus potencias sucesivas

La construcción anterior de **R** recurre de un modo esencial a la distancia que hemos definido entre los números racionales, ya que lo que se considera son sucesiones de números racionales que se van haciendo cada vez más «próximos». Pero también hemos dicho que podían existir varias «distancias» distintas. A principios de siglo se advirtió que se podía dotar al cuerpo **Q** de los racionales de unas distancias llamadas *p*-ádicas.

Un teorema de 1935 debido al matemático Alexander Ostrowski demuestra incluso que la distancia habitual y las distancias p-ádicas son las únicas distancias interesantes con las que se puede dotar al conjunto de los racionales. Si, para completar **Q**, se emplea la distancia *p*-ádica en vez de la distancia usual, se obtiene, no el cuerpo **R** de los reales, sino el «cuerpo de los números *p*-ádicos», representado por **Q***p*.

Ha llegado el momento de explicar la noción de distancia *p*-ádico. Empecemos por el «valor absoluto *p*-ádico» de un entero positivo *n*. *p* representa aquí un cierto número primo como 3, 7, 19, etc., es decir, un entero positivo que sólo es divisible por 1 y por sí mismo. Una vez elegido el valor de *p* se puede intentar dividir *n* por *p* y por sus potencias sucesivas p^2, p^3, p^4, etc. Si la mayor

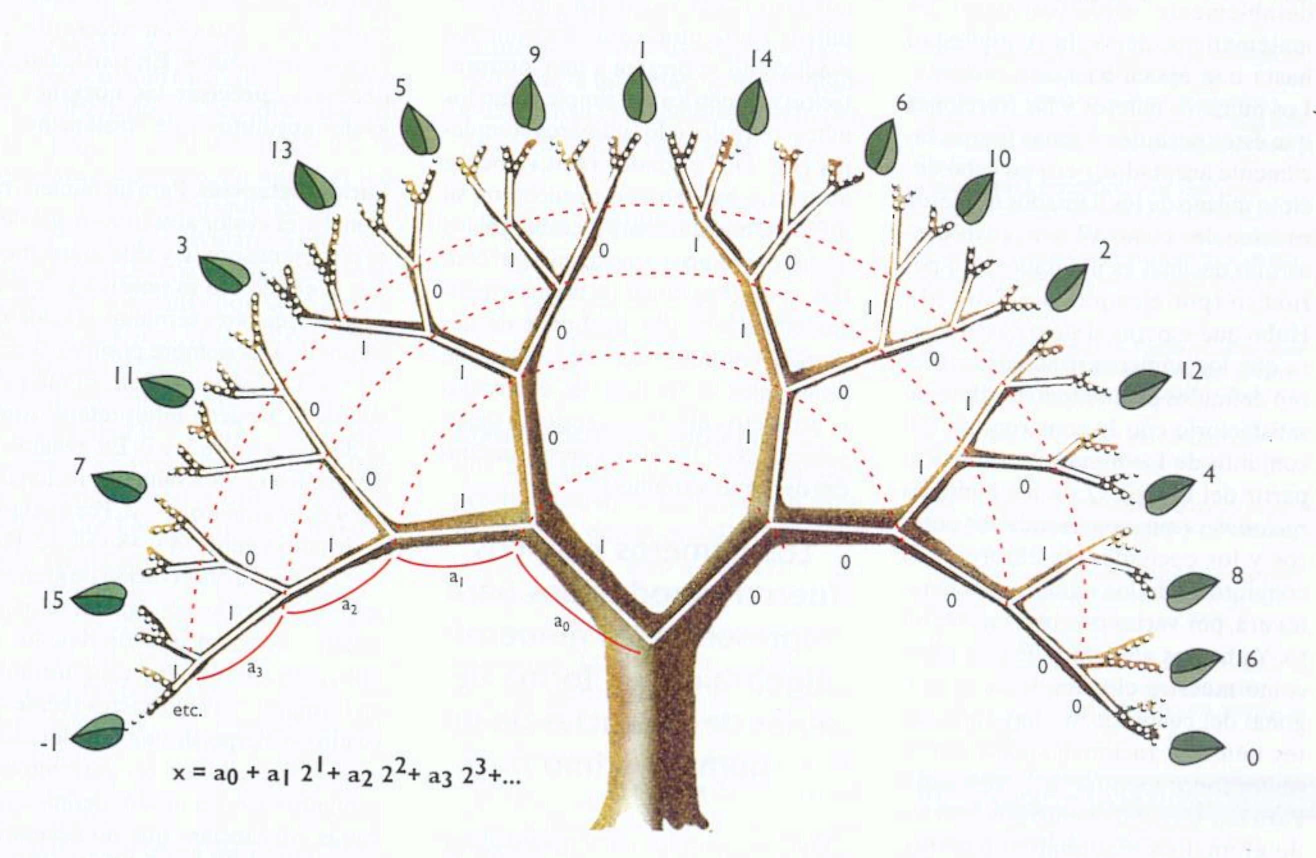


Figura 1. Este árbol infinito en el que cada rama se bifurca representa gráficamente los números 2-ádicos. Más concretamente, el árbol representa los enteros 2-ádicos que se escriben en la forma $a_0 + a_1 2^1 + a_2 2^2 + a_3 2^3 + ...$, donde los coeficientes a_0, a_1, a_2, etc., valen 0 o 1. Se asocia a cada ramificación un coeficiente que vale 0 para la rama de la derecha y 1 para la de la izquierda. Así, cada «hoja» del árbol, obtenida al cabo de una infinidad de ramificaciones, puede ser identificada a un entero «2-ádico». Como los enteros usuales forman parte de los enteros p-ádicos, se han indicado algunos de ellos (por ejemplo, −1, que se escribe $1 + 2^1 + 2^2 + 2^3 + ...$: todos sus coeficientes son iguales a 1).

potencia de p por la cual n es divisible es p^r, entonces el valor absoluto p-ádico de n es, por definición, $1/p^r$. Se escribe entonces $| n |_p = 1/p^r$.

En otros términos, el valor absoluto p-ádico de un entero positivo n es tanto menor cuanto mayor es la divisibilidad de n por p. Por ejemplo, para $p = 5$, el valor absoluto 5-ádico de 26 es $| 26 |_5 = 1$, ya que $26 = 5^0 \times 2 \times 13$; el de 50 vale $| 50 |_5 = 1/25$, pues $50 = 2 \times 5^2$; y el de 375 vale $| 375 |_5 = 1/125$, puesto que $375 = 3 \times 5^3$. Pero si se elige $p = 3$, se tendrá, para los mismos números, $|26 |_3 = | 50 |_3 = 1$ y $| 375 |_3 = 1/3$.

Para un número entero negativo, el valor absoluto p-ádico se define como el de su opuesto. Por ejemplo:

$| -15 |_p = | 15 |_p =$. Por su parte, el valor absoluto p-ádico de 0 es nulo. Para un número racional cualquiera m/n, donde m y n son enteros, se define el valor absoluto p-ádico como el cociente de los valores absolutos del numerador y del denominador: $| m/n |_p = | m |_p / | n |_p =$. Por ejemplo, $| 26/375 |_5 = | 26 |_5 / | 375 |_5 = 1/(1/125) = 125$.

El valor absoluto p-ádico permite definir directamente la distancia p-ádico entre dos números racionales x e y como el valor absoluto p-ádico de su diferencia: $| x - y |_p$. En el caso particular de que x e y sean enteros, la distancia p-ádica es tanto menor cuanto mayor es la potencia de p por la cual es divisible $x - y$. Esta distancia es muy desconcertante comparada con la distancia habitual. En particular, dicha distancia es «ultramétrica», lo cual significa que para x, y, z cualesquiera, la distancia entre x y z es inferior a la mayor de las otras dos distancias, la distancia entre x e y y la distancia entre y y z. Esta curiosa propiedad no se cumple en la geometría usual. No obstante, puede ilustrarse con un árbol genealógico definiendo la distancia entre dos primos como el número de ramas que hay que recorrer en el árbol para pasar de uno a otro a través de un antepasado común (fig. 2). Es fácil constatar que la distancia entre dos primos de la misma generación es como máximo igual a la mayor de las distancias que separan a estos dos primos de un tercero perteneciente a la misma generación.

Una infinidad de potencias. Como hemos dicho ya en varias ocasiones, el conjunto $\mathbf{Q}_p$ de los números p-ádicos se obtiene completando el cuerpo de los racionales por medio de la distancia p-ádica (una vez elegido, claro está, el valor de p). Todo esto, sin duda, parece muy abstracto, pero es posible conseguir una representación más concreta de los números p-ádicos análoga en cierto sentido a la representación decimal de los números reales. En base 10, un número decimal se escribe en la forma: $a_q10^q + ... + a_310^3 + a_210^2 + a_110^1 + a_0 + a_{-1}10^{-1} + a_{-2}10^{-2} + ...$, donde q es un entero positivo y los a_i son enteros comprendidos entre 0 y 9. En la escritura usual, los coeficientes, a_0, a_1, a_2, etc., corresponden a las cifras que van antes de la coma y los coeficientes a_{-1}, a_{-2}, etc. a las que van después. Por ejemplo, la escritura 23,14 equivale al desarrollo $2 \times 10^1 + 3 + 1 \times 10^{-1} + 4 \times 10^{-2}$. Hay un desarrollo análogo para los número p-ádicos, los llamados «desarrollos de Hensel». Se puede demostrar, en efecto, que un número p-ádico puede escribirse siempre bajo la forma $a_{-n}\, p^{-n} + a_{-n+1}\, p^{-n+1} + ... + a_0 + a_1p^1 + a_2p^2 + ...$ donde n es un cierto entero positivo y cada coeficiente entero a_i está comprendido entre 0 y $p - 1$. El desarrollo de Hensel de un número p-ádico puede constar de infinitas potencias de p —o que no significa que los números sean «infinitamente grandes», ya que es la distancia p-ádica la que suministra los criterios de magnitud—. Para los números racionales, que forman parte del conjunto $\mathbf{Q}_p$ de los números p-ádicos, los coeficientes a_i del desarrollo de Hensel se repiten a partir de un cierto punto (de la misma manera que en el cuerpo **R** de los números reales los decimales de un número racional se repiten periódicamente hasta el infinito).

He aquí un ejemplo especialmente simple de desarrollo p-ádico de un número racional. Elijamos $p = 5$ y busquemos el desarrollo de Hensel de $-1/4$. Se escribe dicho número en la forma $1/(1-5)$ y se utiliza la fórmula bien conocida de la serie geométrica $1 + x + x^2 + x^3 + ... = 1/(1-x)$, válida siempre que el valor absoluto de x sea menor que 1. Como el valor absoluto 5-ádico de 5 vale 1/5, que es inferior a 1, la fórmula puede aplicarse. El resultado es $-1/4 =$

Cuerpo, Distancia, Valor Absoluto

CUERPO

Un conjunto K es un «cuerpo» si está provisto de dos operaciones internas, llamadas generalmente suma (+) y producto (×), que verifican las propiedades siguientes:

• K es un grupo conmutativo para la suma:

1) $x + y = y + x$

2) $(x + y) + z = x + (y + z)$

3) Existe un elemento neutro (0) tal que, para todo x, se cumple:
$x + 0 = 0 + x = x$

4) Para todo x, existe un x' tal que $x + x' = 0$

• Asociatividad del producto:
$x \times (y \times z) = (x \times y) \times z$

• Distributividad del producto respecto de la suma:
$x \times (y + z) = x \times y + x \times z$ y
$(x + y) \times z = x \times z + y \times z$

• Existe un elemento neutro $e \neq 0$ para el producto: $x \times e = e \times x = x$

• Todo $x \neq 0$ posee un inverso x'', es decir, $x \times x'' = x'' \times x = e$.

Si el producto es conmutativo ($x \times y = y \times x$) se dice que K es un cuerpo conmutativo. El conjunto conjunto **Q** de los números racionales, el conjunto R de los números reales, el conjunto $\mathbf{Q}_p$ de los números p-ádicos y el conjunto C de los números complejos son cuerpos conmutativos.

DISTANCIA

Una distancia en un conjunto E es una aplicación d que asocia a todo par (x, y) de elementos de E un número real positivo o nulo $d(x, y)$ que verifica:

• $d(x, y) = d(y, x)$

• $d(x, y) > 0$ si $x \neq y$ y $d(x, x) = 0$

• $d(x, z) \leq d(y, x) + d(y, z)$

Se dice que la distancia es ultramétrica cuando $d(x, z) \leq$ máx $d(x,y), d(y, z))$. Las distancias p-ádicas son ultramétricas.

VALOR ABSOLUTO

Un valor absoluto sobre un cuerpo K es una aplicación que a todo elemento x de K le hace corresponder un número real positivo o nulo, representado por $| x |$, tal que:

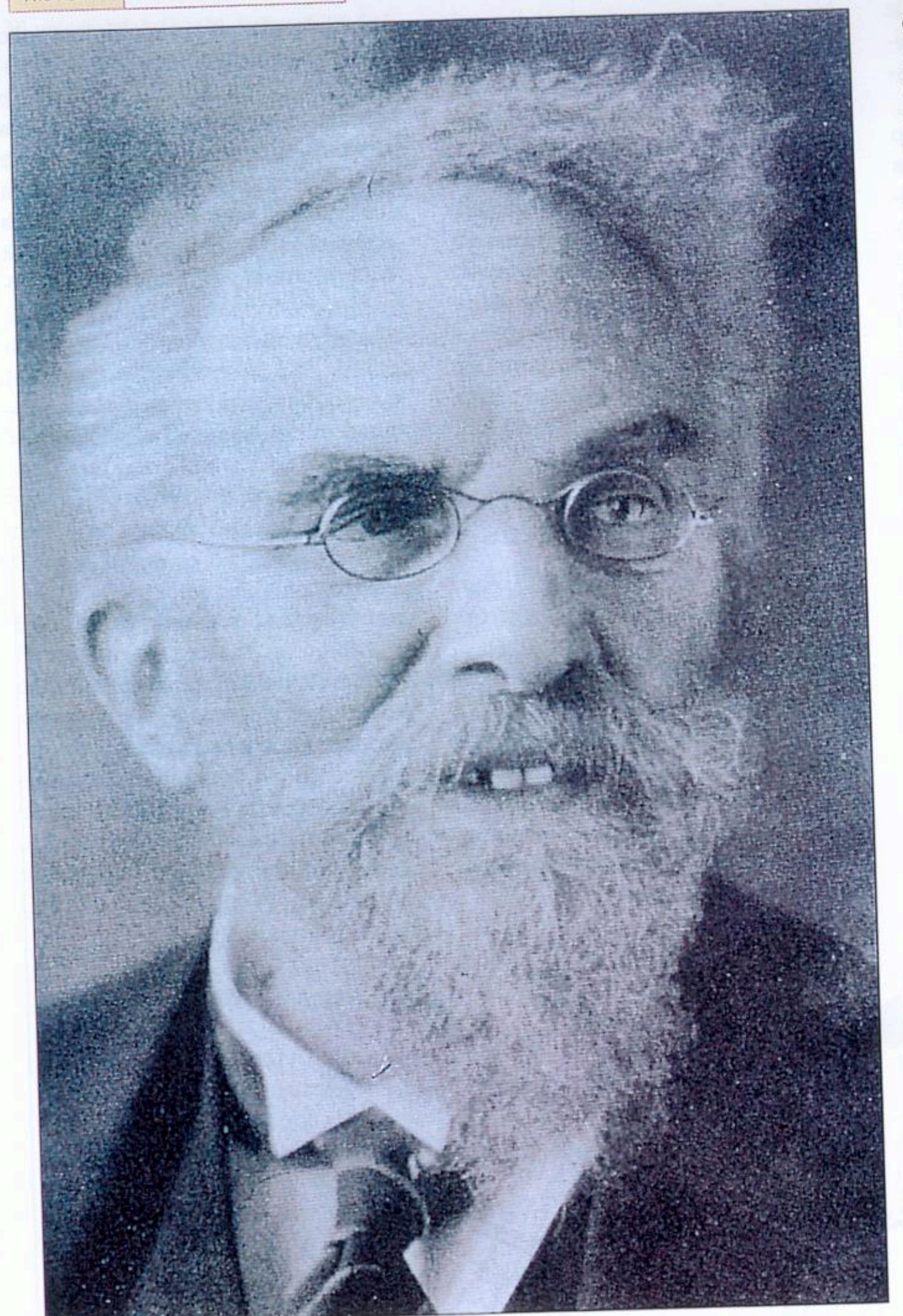
• $| x | > 0$ si $x \neq 0$ y $| 0 | = 0$.

• $| x + y | \leq | x | + | y |$

(para un valor absoluto ultramétrico, $| x + y | \leq$ máx $(| x |, | y |)$)

• $| x \times y | = (| x | \times | y |$

En tales condiciones, la expresión $d(x, y) = | x - y |$ define una distancia sobre el cuerpo K.



El matemático alemán Kurt Hensel inventó los números p-ádicos a principios del siglo XX. Hensel, que fue alumno del famoso teórico de números Lepold Kronecker, enseñó en Berlín y en la Universidad de Marburgo. (Foto Jean-Loup Charmet)

$1 + 5 + 5^2 + 5^3 + ...$, que es el desarrollo de Hensel de –1/4. Los coeficientes, a partir de a_0, son todos iguales a 1. Pero lo que antecede no es más que un ejemplo y el desarrollo de Hensel no suele ser tan fácil de obtener.

Este tipo de desarrollo, dicho sea de paso, permite representar geométricamente los números *p*-ádicos. Imaginemos un árbol cada una de cuyas ramas se ramifica en *p* ramas secundarias, que se subdividen a su vez y así hasta el infinito. Numerando 0, 1, 2, ... *p* –1 las ramas que nacen de cada nudo, se puede representar cada coeficiente a_n del desarrollo de Hensel de un número *p*-ádico por medio de una de las ramas del árbol, la cual a su vez procede de la rama madre correspondiente al coeficiente a_{n-1}. Se asimila así cada uno de los números *p*-ádicos a una «hoja» del árbol, que es un extremo de esta ramificación hasta el infinito. El ejemplo más fácil de dibujar es el del cuerpo $\mathbf{Q}_2$ de los números 2-ádicos (fig. 1). Se parte de una rama madre que se divide en dos ramas hijas. Se decide, por ejemplo, que la rama de la derecha corresponde al valor 0 y la de la izquierda al valor 1 (los dos valores posibles del coeficiente a_0) y se repite el proceso hasta el infinito. Dado que un número 2-ádico está definido por una sucesión infinita de 0 y 1 (los valores de los coeficientes del desarrollo de Hensel), puede identificarse con un determinado camino a través del árbol o, lo que es lo mismo, con una de las «hojas» que culminan la ramificación infinita.

Como indica todo lo que antecede, el manejo de los números *p*-ádicos no es tan cómodo como el de los números reales. No obstante, una vez superadas estas dificultades iniciales, es posible emprender un análisis *p*-ádico, es decir, examinar objetos matemáticos clásicos, como sucesiones, series, funciones, ecuaciones algebraicas, ecuaciones diferenciales, etc., situándose en el universo de los números *p*-ádicos y no en el de los números reales. Algunos resultados son más simples que en el análisis clásico (por ejemplo, toda serie $u_1 + u_2 + ... + u_n$ cuyo término general tiende a 0 cuando *n* tiende a infinito converge hacia un cierto límite). Pero también aparecen ciertos fenómenos nuevos, los cuales, por supuesto, son los más interesantes de estudiar.

> Debido a su estructura arborescente, estos números están invadiendo poco a poco las probabilidades y la física teórica

Dos modos de utilización. ¿Para qué sirven los números *p*-ádicos y el análisis que permiten elaborar? Para los matemáticos al menos, para muchas cosas. Aquí sólo podemos esbozar un breve e incompleto resumen. En primer lugar, una observación: hay dos maneras de utilizar los números *p*-ádicos. O bien considerando una sola distancia *p*-ádica, en cuyo caso el número primo p elegido desempeña un papel privilegiado, o bien haciendo intervenir simultáneamente todas las distancias que es posible definir en los números racionales (tanto las distancias *p*-ádicas como la distancia clásica). El interés de este segundo enfoque reside en la llamada «fórmula del producto» que relaciona todas estas distancias: el producto de todos los valores absolutos *p*-ádicos de un número racional m/n es

igual al inverso del valor absoluto ordinario de m/n (fórmula que se demuestra fácilmente descomponiendo m y n en factores primos).
Combinando estos dos enfoques, se pueden obtener unos resultados que se expresan de manera clásica, es decir, en cuyo enunciado no aparecen los números p-ádicos. Por ejemplo, métodos de esta clase permitieron en 1990 al neerlandés Fritz Beukers y a los franceses Jean-Paul Bezivin y Philippe Roba obtener una nueva demostración del teorema de Lindemann-Weierstrass (que implica, entre otras cosas, la trascendencia de π).
Hace muy poco, el francés Yves André ha generalizado considerablemente estas técnicas mostrando cómo utilizarlas para demostrar la irracionalidad o la trascendencia de valores de funciones solución de ecuaciones diferenciales de tipos bastante generales. Parece el punto de partida de un estudio aritmético de las ecuaciones diferenciales.

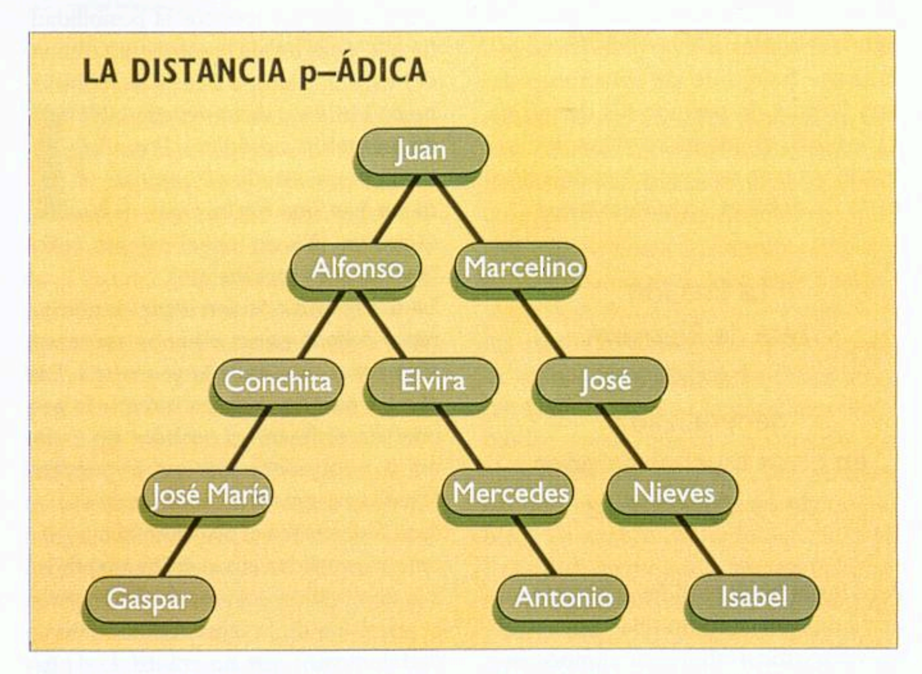


Figura 2. Una distancia p-ádica es ultramétrica, lo que significa que para x, y, z cualesquiera, la distancia $d(x, y)$ entre x e y es menor o igual que la mayor de las distancias $d(x, z)$ y $d(z, y)$. Esta propiedad puede ilustrarse por medio de un árbol genealógico contando el número de ramas que separa a dos primos de la misma generación y considerando este número como la distancia entre ambos. Aquí, por ejemplo, la distancia entre Gaspar y Antonio vale 6 y la distancia entre Antonio e Isabel vale 8. Por lo tanto la distancia entre Gaspar e Isabel debe ser menor o igual que la mayor de las otras dos (8), lo que puede comprobarse sin ninguna dificultad.

Teorema de Fermat. Pero el ejemplo más espectacular de la utilidad de los métodos p-ádicos lo dio recientemente el inglés Andrew Wiles en su demostración del teorema de Fermat (véase el artículo de C. Goldstein en este número). Este teorema dice que que para n mayor que 2 no existen enteros positivos y no nulos a, b y c tales que $a^n + b^n = c^n$. La demostración hace un uso esencial, en varias ocasiones, de los números p-ádicos, mientras que el enunciado versa sobre números ordinarios y es muy anterior a la invención de los p-ádicos.
La ecuación de Fermat forma parte de las llamadas «ecuaciones diofánticas», cuyo estudio constituye una de las primeras aplicaciones de los números p-ádicos. Las ecuaciones diofánticas, así llamadas en honor del matemático griego Diofanto, que estudió algunas en el siglo IV, son ecuaciones polinómicas de una o varias variables con coeficientes enteros (por ejemplo, $(3x^4 + 5y^6 - 2xy = 0)$ que aparecen a menudo en matemáticas y en las aplicaciones de esta ciencia. Muchas ecuaciones polinómicas carecen de soluciones reales. Es el caso de la ecuación $x^2 + 1 = 0$. Para resolverla, los matemáticos del siglo XVI inventaron los números complejos, basados en la introducción del símbolo $\sqrt{-1}$. Pero hay otra posibilidad que recurre a los números p-ádicos: se puede demostrar que si si p es un número primo de la forma $4n + 1$, existe un número p-ádico x tal que $x^2 + 1 = 0$.

Eligiendo $p = 5$, los cálculos demuestran que su desarrollo de Hensel comienza por $2 + 5 + 2 \times 5^2 + 5^3 + 3 \times 5^4 + 4 \times 5^5 + ...$ Este ejemplo ilustra el hecho de que una ecuación sin solución real puede poseer una solución p-ádica. Generalmente, cuando se habla de ecuaciones diofánticas, se sobreentiende que se buscan soluciones enteras. ¿En qué medida contribuyen los números p-ádicos a resolver este tipo de problemas? Para averiguarlo, baste el comentario siguiente: si una ecuación carece de solución real, carece *a fortiori* de solución entera, pues los enteros forman parte de los reales. Por ejemplo, para que un número entero sea solución del trinomio $3x^2 + 2x - 1 = 0$, es necesario (pero no suficiente) que dicha ecuación admita una solución real x. Este criterio elemental puede trasponerse directamente a los números p-ádicos, que incluyen también a los enteros: para que una ecuación diofántica admita una solución entera, es necesario —pero no suficiente— que admita una solución p-ádica para todo p primo.
Resulta que se dispone de potentes métodos para estudiar las ecuaciones en los cuerpos p-ádicos. He aquí un ejemplo, que forma parte de los llamados «lemas de Hensel». Sea un polinomio $P(x)$ de coeficientes enteros. Supongamos que se conoce un entero n tal que el entero $P(n)$ es divisible por p pero $P'(n)$ no lo es (donde $P'(x)$ es la derivada de $P(x)$). Existe entonces un número p-ádico tal que $P(x) = 0$. Por ejemplo, si $P(x) = x^2 + 1$, su derivada es $P'(x) = 2x$; se cumple, pues, que $P(2) = 5$ y $P'(2) = 4$. Como el número 4 no es divisible por 5, el polinomio $x^2 + 1$ admite una raíz en el conjunto Q_5 de los números 5-ádicos. Hemos dado más arriba los primeros términos del desarrollo de Hensel de esta raíz.

Un problema abierto. Por medio de técnicas como la anterior se consigue eventualmente hallar un número primo p para el cual la ecuación carece de raíces en los números p-ádicos. Se concluye, pues, que la ecuación carece de solución entera. Otro resultado importante, llamado «principio de Hasse» en honor de Helmut Hasse, alumno de Hensel, enuncia el recíproco para las ecuaciones diofánticas de grado 2 y una o varias variables. Si una tal ecuación admite una solución p-ádica para todo número primo p y una solución

La función zeta de Riemann, una función notable por los resultados puramente aritméticos que permite obtener.

real, entonces admite una solución entera. Por desgracia, este principio carece de validez general. El hallazgo de las condiciones de validez del principio de Hasse es un problema muy estudiado pero todavía en gran medida abierto.

El empleo de los números p-ádicos ha sido extendido también a las funciones, especialmente a la «función zeta de Riemann»,una función notable por los resultados puramente aritméticos que permite obtener. Esta función $\zeta(s)$ se define así: $\zeta(s) = 1 + 1/2^s + 1/3^s + 1/4^s + ...$ para $s > 1$. Como demostró Bernhard Riemann en el siglo XIX, no puede prolongarse para ningún valor complejo de la variable s (excepto para $s = 1$). La función así construida está estrechamente ligada a las propiedades de los números primos (véase el artículo de Henri Cohen en este mismo número). La función zeta de Riemann ha sido generalizada en otros muchos campos de la matemática. En particular, también en el siglo pasado, el alemán Peter Gustav Lejeune-Dirichlet, en sus investigaciones sobre los números primos contenidos en una progresión aritmética, introdujo unas funciones de variable compleja llamadas actualmente «funciones L de Dirichlet», muy útiles en aritmética.

En 1964, los matemáticos Tomio Kubota y Heinrich W. Leopold consiguieron crear las técnicas necesarias para construir análogos p-ádicos de las funciones zeta y L. Estos nuevos objetos, que son funciones de una variable p-ádica que toma valores en los números p-ádicos, permiten obtener valiosas informaciones de tipo aritmético complementarias de las que suministran las funciones zeta y L clásicas, pero que por desgracia no podemos exponer aquí. En cualquier caso, a partir de los años 1960, gracias a los trabajos del estadounidense Bernard Dwork sobre las funciones zeta asociadas a «variedades algebraicas» (conjunto de soluciones de una familia de polinomios de varias variables), el análisis p-ádico ha cobrado un gran empuje y ha adquirido carta de nobleza en matemáticas.

La función zeta de Riemann ha sido generalizada en otros muchos campos de la matemática

Pero los métodos p-ádicos no sólo intervienen en matemáticas puras. Se les ve aparecer ahora en campos inesperados, como las probabilidades y la física teórica. Una de las razones de esa presencia es que los números p-ádicos proporcionan un ejemplo sencillo de estructura arborescente. Por ejemplo, en el estudio teórico de las propiedades de los vidrios de espín (materiales desordenados que contienen partículas imanadas cuya orientación debe ajustarse para minimizar las interacciones magnéticas), la llamada «técnica de las réplicas» consiste en considerar n muestras idénticas, calcular la energía de interacción magnética y hacer tender formalmente, en los cálculos, el entero n a cero. Un examen atento pone de manifiesto que la técnica de las réplicas equivale a considerar una serie de enteros que tiende «p-ádicamente» a cero para todos los números primos p a la vez (por ejemplo, la sucesión $n \rightarrow n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$ posee esta propiedad porque, para todo número primo p, $n!$ es divisible por potencias de p cada vez mayores al tender n a infinito).

Estructura del espacio-tiempo. Las aplicaciones del análisis p-ádico a la física podrían incluso ir más allá de los aspectos estrictamente técnicos. Por ejemplo, los físicos teóricos especulan sobre la estructura del espacio y del tiempo a muy pequeña escala. Las leyes de la relatividad y de la física cuántica parecen indicar que no es posible medir longitudes inferiores a un valor extraordinariamente pequeño, llamado longitud de Planck, del orden de 10^{-35} metros. La existencia de una distancia mínima sugiere a muchos teóricos la posibilidad de que a esa escala la estructura última del espacio-tiempo pueda describirse no en términos de números reales sino de estructuras p-ádicas. Por ahora se trata sólo de estudios especulativos, pero no hay que excluir que desemboquen un día en conclusiones verificables por experimentos.

La situación a que dan lugar los números p-ádicos es en algunos aspectos muy parecida a la de la geometría. En el siglo pasado se descubrió que la geometría ordinaria o euclídea no es la única a considerar, ya que se pueden construir distintas geometrías no euclídeas. Se pensaba al principio que la geometría euclídea era la única adaptada a la descripción del mundo físico, pero la aparición de la teoría de la relatividad demostró que no era así. Los números p-ádicos han sido y siguen siendo una fuente de grandes progresos en aritmética y geometría algebraica. ■

DANIEL BARSKY es director de investigación del CNRS y trabaja en la Universidad Paris-Nord (Villetaneuse). Se interesa por las funciones *L* *p*-ádicas.

GILLES CHRISTOL es investigador de la Universidad de Paris-VI e investiga sobre ecuaciones diferenciales en cuerpos *p*-ádicos.

PARA MÁS INFORMACIÓN:

- G. Christol, «*p-adic Numbers and Ultrametricity*» en *From Number Theory to Physics* (Waldschmidt *et al.*, eds.), Springer-Verlag, 1992.
- R. Rammal *et al.*, «Ultrametricity for Physicists», *Rev. Mod. Phys.*, *58*, 1986.
- Y. Amice, **Les Nombres p-adiques**, PUF, 1975.
- N. Koblitz, *p-adic Numbers, p-adic Analysis and Zeta Functions*, 2nd ed.,Springer-Verlag, 1984.
- J.A. Paulus, *Más allá de los números*, Tusquets Editores, Colección Metatemas, Barcelona, 1993.

Veintitrés problemas. Una manera de orientar nuestra reflexión sobre este tema consiste en recordar la célebre lista de veintitrés problemas que el matemático alemán David Hilbert propuso en 1900 como reto al siglo XX naciente (véase el artículo de J.-P. Bourguigon y H. Sinaceur en este número). Uno de los problemas, el sexto de la lista, consistía en axiomatizar la física, es decir, en expresar todas las leyes fundamentales de la física en forma de reglas matemáticas formales; esta cuestión englobaba la axiomatización de la teoría de probabilidades, pues para Hilbert las probabilidades tenían que ver con el mundo real y eran de la incumbencia de la física. Su décimo problema, por su parte, concernía a las ecuaciones «diofánticas», es decir, a las ecuaciones algebraicas en las que se buscan soluciones en forma de números enteros. La pregunta de Hilbert era: «*¿Hay algún medio de decidir si una ecuación algebraica posee solución en términos de números enteros?*» Hilbert estaba lejos de sospechar la existencia de una relación entre ambos problemas, una relación que, como veremos, existe sin lugar a dudas.

Para Hilbert y para la mayoría de los matemáticos de la época, la idea de que todo problema matemático posee una solución era algo que caía por su propio peso. Sólo más tarde reconoció Hilbert que había allí un tema que explorar. De esta exploración ha resultado que una pregunta matemática clara y simple no siempre posee una respuesta inequívoca; además, una cierta forma de azar aparece incluso en matemáticas puras y se encuentra en las ecuaciones diofánticos, objetos del décimo problema de Hilbert. Veremos, en efecto, que ciertas cuestiones bastante simples de aritmética, ligadas a las ecuaciones diofánticas, tienen —en un sentido bien determinado— una respuesta completamente aleatoria. Y ello, no porque no podamos resolverlas mañana, sino porque la respuesta es aleatoria cualquiera que sea el razonamiento utilizado.

¿Cómo se ha llegado aquí? Un primer punto tiene que ver con la noción de razonamiento axiomático, es decir, de razonamiento matemático basado en reglas formales. El sistema geométrico de Euclides es un ejemplo simple de sistema axiomático, pero desde fines del siglo XIX se han propuesto varios sistemas de axiomas para formalizar completamente las matemáticas, así como la lógica en la que se basa todo razonamiento humano. La axiomática y el fundamento de las matemáticas han sido estudiados por muchos investigadores, incluido el propio Hilbert. Este último, en particular, había formulado una exigencia: para que un sistema de axiomas sea satisfactorio tiene que existir un «*procedimiento mecánico*», es decir, una sucesión de operaciones lógicas en número finito que permita decidir si una demostración matemática cualquiera verifica o no las reglas formales fijadas. Se trata de una exigencia de claridad y objetividad que parece perfectamente natural. Lo importante para lo que sigue es que si se construye un sistema de axiomas *coherente* (es decir, tal que en él no pue-



dan demostrarse simultáneamente un resultado y su contrario) y *completo* (es decir, tal que toda aserción sea en él verdadera o falsa) entonces se sigue inmediatamente que existe un procedimiento mecánico que en principio permite zanjar cualquier pregunta que pueda formularse en el marco de dicha teoría.

¿Se detendrá el programa al cabo de un tiempo finito? Razonando por el absurdo se puede demostrar que el problema es insoluble

Un gigantesco inventario. Un tal procedimiento consistiría (al menos en principio, pues en la práctica el tiempo necesario sería prohibitivo) en listar todas las demostraciones posibles escritas en el lenguaje formal, es decir, en el sistema de axiomas elegido, por orden de tamaño y por orden alfabético de los símbolos empleados. Es lo que se puede designar figuradamente como el *«algoritmo del British Museum»* para aludir al gigantismo del «inventario» a efectuar. En otros términos, se enumeran todas las demostraciones posibles y se verifica si derivan de las reglas formales del sistema axiomático. Se obtienen así en principio todos los teoremas, todo lo que puede demostrarse en el marco del sistema de axiomas. Y si éste es coherente y completo, toda afirmación podrá ser confirmada (si está demostrada) o refutada (si está demostrada su contraria). Se obtiene así un procedimiento mecánico que permite decidir si una aserción es verdadera o falsa.

Por desgracia, la situación no resultó ser tan simple. Desde los trabajos fundamentales del austríaco Kurt Gödel en 1931 y del británico Alan M. Turing en 1936, sabemos que la empresa es vana: no existe ningún sistema axiomático coherente y completo para la aritmética y además no puede haber ningún procedimiento mecánico que permita determinar, para toda aserción matemática, si es verdadera o falsa.

De este resultado que ha marcado profundamente el pensamiento matemático dio Gödel una demostración muy ingeniosa: es su célebre *«teorema de incompletitud»*. Pero el modo de proceder de Turing me parece en cierto modo más fundamental y más fácil de comprender. Me refiero aquí al teorema de Turing, según el cual no existe ningún procedimiento matemático capaz de determinar, para un programa informático arbitrario, si se ejecutará o no en un tiempo finito una vez puesto en marcha. De ahí se sigue inmediatamente el teorema de Gödel: si no hay ningún procedimiento mecánico para determinar si un programa se detiene o no en un tiempo finito, entonces tampoco puede haber un sistema de axiomas capaz de hacerlo.

Sin entrar en detalles, se puede esbozar una manera de demostrar que el problema de la detención de un programa es insoluble. Se trata de una demostración por el absurdo: Supongamos que existe un procedimiento mecánico que permita averiguar, para todo programa, si se ejecutará en un tiempo finito. Esto implica la posibilidad de construir un programa (P) que incorpore el dato de un número entero N y efectúe las tareas siguientes: primero, examinar todos los programas posibles de tamaño menor o igual a N bits (todo programa informático puede traducirse a una sucesión de cifras binarias, 0 o 1, llamadas bits, cada una de las cuales constituye una unidad de «información») y determinar cuáles se detienen en un tiempo finito. Luego, simular la ejecución de la totalidad de estos últimos y considerar sus resultados. Supongamos que los resultados son números enteros positivos o nulos, lo cual cabe hacer sin pérdida de generalidad, ya que todo programa produce como resultado una sucesión de 0 y 1, y ésta puede interpretarse siempre como representación de un número entero positivo o nulo. La última tarea que se asigna al programa (P) consiste en tomar el resultado máximo producido por todos los programas que se detienen en un tiempo finito y cuyo tamaño no rebasa los N bits y calcular el doble (por ejemplo) de este resultado máximo.

Principal ingrediente. Examinemos ahora la situación a la que se llega. El número N es la información básica incluida en el programa (P) que acabamos de describir. Por tanto, el tamaño de este programa es del orden de $\log_2$ N bits, pues para expresar el número N se requieren $\log_2$ N bits en el sistema binario (por ejemplo, el número 109 se escribe 110110 en el sistema binario, lo cual requiere $7 \approx \log_2 109$bits). Por supuesto, el programa P debe contener también otras instrucciones que permitan enumerar y simular la ejecución de todos los programas de tamaño in-





El lógico austríaco Kurt Gödel (a la izquierda) socavó en 1931 con la convicción íntima de casi todos los matemáticos de que es posible construir unos sistemas formales de axiomas que sean completos y coherentes. Gödel demostró que todo sistema formal contiene enunciados indecidibles, es decir, que no pueden ser confirmados o refutados sobre la única base de los axiomas del sistema. El británico Alan M. Turing (a la derecha) formalizó las nociones que son los fundamentos teóricos de la informática. En particular demostró en 1936 que no existe ningún procedimiento que permita averiguar si un programa arbitario se ejecutará en un tiempo finito o no. (Foto AFP y ET F.T. archive)

¿Forman los decimales de π una sucesión aleatoria?

¿En qué sentido puede calificarse de aleatoria la sucesión de las cifras que componen un número? La cuestión no es tan simple como parece. Hace aproximadamente un siglo, el matemático francés Emile Borel (1871-1956)había definido en este contexto la noción de número «normal» y había demostrado que casi todos los números son normales.

¿Qué es un número normal? Un número se llama normal en una base b si en el desarrollo del número según esta base cada una de las b cifras posibles aparece con la misma frecuencia 1/b, si cada uno de los b^2 grupos de 2 cifras aparece con la misma frecuencia $1/b^2$, y lo mismo con los grupos de tres, cuatro cifras, etc. Por ejemplo, un número es normal en el sistema binario (b = 2) si en su desarrollo binario las cifras 0 y 1 aparecen con la misma frecuencia límite 1/2, si las secuencias 00, 01, 10 y 11 aparecen con la misma frecuencia 1/4, etc.

Un número se llama *absolutamente* normal si es normal cualquiera que sea la base b en la que se expresa. E. Borel demostró en 1909 que *casi todos* (expresión que tiene un sentido matemático preciso) los números reales son absolutamente normales.



En otros términos, elijamos un número echando a cara o cruz cada uno de los bits que constituyen su desarrollo infinito en el sistema binario. Entonces, el número comprendido entre 0 y 1 elegido de esta manera es absolutamente normal y ello «casi con seguridad», es decir, con probabilidad igual a 1.

Si la no-normalidad es la excepción a la regla, se podría pensar que es fácil hallar ejemplos de números normales. ¿Qué decir, por ejemplo, de $\sqrt{2}$, π o e? ¿Son normales? Curiosamente, no lo sabemos. Se han hecho numerosos cálculos por ordenador para obtener las cifras sucesivas de estos números y determinar su frecuencia de aparición. Todo ocurre como si fueran normales pero nadie hasta hoy ha podido demostrarlo rigurosamente. De hecho, ha sido extraordinariamente difícil exhibir un solo ejemplo de número normal. En 1933, D.G. Champernowne logró dar con un número del que demostró que era normal en el sistema decimal; este número se escribe: 0,0 1 2 3 4 5 6 7 8 9 10 11 12... 98 99 100 101 102... 998 999 1000 1001 1002...

Pero no se sabe si este número es absolutamente normal, es decir, normal en toda base.

Sin embargo, disponemos actualmente de un ejemplo natural de número absolutamente normal: la probabilidad de detención Ω de la que se ocupa el artículo. En efecto, se puede demostrar fácilmente que Ω es absolutamente normal a partir del hecho de ser algorítmicamente aleatorio. Un número es algorítmicamente aleatorio si, para determinar N bits de su desarrollo binario, hace falta un programa cuyo tamaño sea de al menos N bits. Para dar un contraejemplo, los N decimales de los números 0,111111111... y 0,11011011010 pueden calcularse muy fácilmente por medio de un programa cuyo tamaño es muy inferior a N (para N no demasiado pequeño). En efecto, basta traducir las órdenes «repetir N veces la cifra 1» o «repetir N veces la secuencia 001» en lenguaje binario. Estos números, por tanto, no son en absoluto algorítmicamente aleatorios.

Aunque, $\sqrt{2}$, π o e sean normales (lo que está por demostrar), no pueden ser algorítmicamente aleatorios porque existen algoritmos de tamaño finito para calcular sus cifras sucesivas. El número de Champernowne es incluso peor en este sentido: no sólo sus cifras son calculables y previsibles, sino que es muy fácil hacerlo. Como se ve, la noción de número algorítmicamente aleatorio es mucho más fuerte que la de número normal. De la misma manera, la gran importancia práctica de los algoritmos que producen números pseudoaleatorios (utilizados en los juegos informáticos o en ciertos métodos de cálculo numérico) reside precisamente en hecho de que las sucesiones de números producidas son extremadamente compresibles algorítmicamente hablando.

ferior a N bits, pero el resultado no quedará sustancialmente modificado: el programa (P) tiene efectivamente un tamaño del orden de $\log_2 N$ bits (y por tanto inferior a N bits). Este punto requiere tal vez alguna aclaración. Ingenuamente, se tiende a pensar que (P) debe contener todos los programas de menos de N bits. Pero no por simular su ejecución debe contenerlos. Para dar una ilustración, un programa encargado de efectuar la suma de todos los enteros comprendidos entre 1 y 1000 no necesita tener en memoria todos los enteros de 1 a 1000: los va produciendo a medida que va realizando los cálculos de la suma. Baste esto para indicar que N es realmente el ingrediente principal del programa (P). Pero retomemos el hilo: por construcción, este programa produce un resultado que es como mínimo el doble de grande que el que produce todo programa de tamaño inferior a N bits. Pero esto es contradictorio, ya que (P) forma parte de estos programas, por lo que daría un resultado al menos dos veces mayor que el que daría... La hipótesis de partida (la existencia de (P)) es falsa. Por tanto, el problema de la detención de un programa es insoluble, lo cual acabamos de demostrar utilizando el punto de vista de la teoría de la información.

Juego de palabras matemático. Partamos de este resultado fundamental de Turing para obtener mi resultado de 1987[1] sobre el azar en matemáticas; basta modificar el vocabulario. Es una

especie de juego de palabras matemático. De la insolubilidad del problema de la detención se pasa al azar ligado a la *probabilidad* de detención. ¿Qué es esta última? An vez de preguntarse si un determinado programa se detendrá o no al cabo de un tiempo finito, se considera el conjunto de todos los programas informáticos posibles, lo cual puede hacerse en principio con un ordenador idealizado que en la jerga matemática se llama *«calculador universal de Turing.»* A cada programa posible se le asocia una probabilidad (que no hay que confundir con la probabilidad de detención, que pronto definiremos). Como todo programa, en definitiva, es equivalente a una sucesión de bits, se elige cada bit al azar, por ejemplo jugándolo a cara o cruz: a un programa de N bits se le asociará pues la probabilidad $1/2^N$. De hecho, nos limitamos a los programas bien estructurados, de los que suponemos que terminan con la instrucción *«fin de programa»*, la cual no puede aparecer al comienzo o a la mitad del programa; en otros términos, ningún programa bien estructurado constituye la extensión de otro programa bien estructurado. Esta hipótesis es técnica pero esencial, pues en su ausencia la probabilidad total $1/2^N$ sería mayor que 1 (e incluso infinita). Se define entonces la probabilidad de detención Ω (omega) como la probabilidad de que, habiendo elegido al azar un programa, éste se ejecute en un número finito de etapas. Este número Ω vale $\Sigma_N(a_N/2^N)$, donde a_N es el número de programas bien estructurados de N bits que se ejecutan en un tiempo finito.

Ω es una probabilidad, y por tanto un número comprendido entre 0 y 1. Si se obtuviera Ω = 0, esto significaría que ningún programa se detiene y si se obtuviera Ω = 1, que se detienen todos. Esta probabilidad puede expresarse en distintas bases; una base particularmente conveniente es la base binaria, en la cual el número Ω es una sucesión de 0 y 1; por ejemplo 0,111010001101.

Edouard A. Lucas demostró el teorema que iba a resolver el 10º problema de Hilbert sobre las ecuaciones diofánticas

Información irreductible. La cuestión que cabe plantear entonces es la siguiente «¿Cuál es el N-ésimo bit de la probabilidad de detención Ω?»

La aserción de Turing (*«el problema de la detención es indecidible»*) lleva a mi resultado de que la probabilidad de detención es aleatoria, o más exactamente de que constituye una *información matemática irreductible*. En otros términos, cada bit de la representación binaria de Ω es un hecho matemático que es lógicamente y estadísticamente independiente de los demás. Saber si un determinado bit de Ω es un 0 o un 1 es un hecho matemáticamente irreductible que no puede condensarse o reducirse más. Una manera más precisa de decir lo mismo es que la probabilidad de detención es *algorítmicamente aleatoria*, es decir, que para calcular N bits de la representación binaria de π hace falta un programa informático cuyo tamaño es de al menos N bits (véase el recuadro «*¿Forman los decimales de _ una sucesión aleatoria?*»). Una manera resumida de expresar lo mismo es la siguiente: «La aserción de que el N-ésimo bit de Ω es 0 o 1, para un N dado, es un hecho matemático aleatorio análogo al resultado de echar un moneda al aire.»

Un problema insoluble. Se replicará inmediatamente que este no es el tipo de afirmaciones que se encuentran habitualmente en matemáticas puras. Nos gustaría poder traducir este enunciado al lenguaje de la teoría de números, que constituye los cimientos de las matemáticas. Gödel tuvo que vérselas con el mismo problema. La afirmación verdadera pero indemostrable que había construido era extraña; decía. «Soy indemostrable». Gödel desplegó grandes dosis de ingenio y utilizó unos razonamientos muy sofisticados para transformar «Soy indemostrable» en un enunciado sobre números enteros. Los trabajos de Gödel han dado origen a numerosas investigaciones cuya conclusión final es que el décimo problema de Hilbert es insoluble: no hay ningún algoritmo que permita determinar en un número finito de operaciones si una ecuación idofántica arbitraria posee solución.Este problema resulta equivalente al de Turing sobre la detención de un programa: dado un programa informático, cabe construir una ecuación diofántica que posea una solución si y sólo si este programa se ejecuta en un tiempo finito. Recíprocamente, dada una ecuación diofántica, se puede construir un programa que se detenga si y sólo si esta ecuación posee solución.

Especialmente espectaculares son en este contexto los trabajos de los matemáticos James P. Jones, de la Universidad de Calgary en Canadá, y Yuri V. Matijasevic, del Instituto Steklov de San Petersburo, antiguamente Leningrado, publicados hace unos quince años.[2]

Estos dos matemáticos observaron que existía un teorema muy simple, demos-

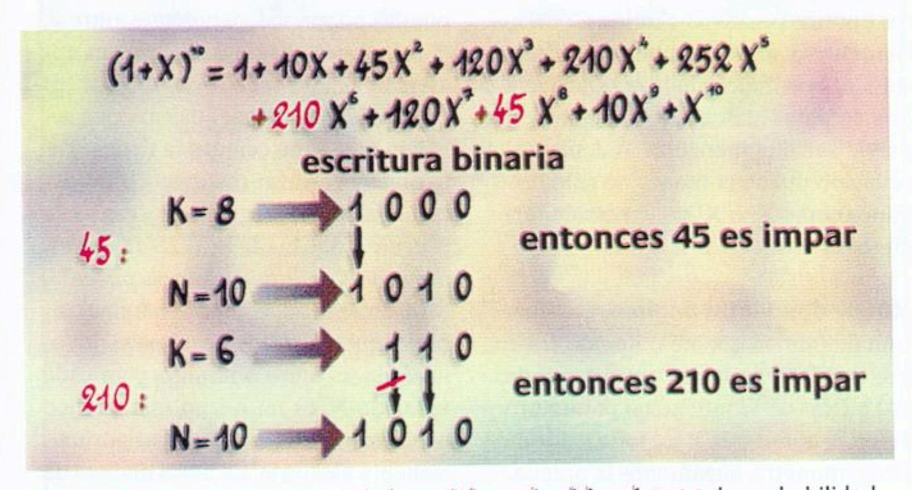


Fig. 1. Dado un programa informático elegido al azar, la probabilidad Ω de que se ejecute en un tiempo finito puede escribirse en un sistema binario en forma de una sucesión de 0 o 1 llamados bits. Para obtener una ecuación que determine los bits de la probabilidad de detención de un programa elegido al azar, el autor ha utilizado unas técnicas basadas en un teorema simple debido a un matemático francés del siglo pasado, Edouard A. Lucas. Este teorema afirma que el coeficiente de X^K en el desarrollo de $(1 + X)^N$ es impar si y sólo si los «1» de las escritura binaria de K están en el mismo lugar en la escritura binaria de N.

JUGAR A CARA O CRUZ CON UN ECUCACIÓN DIOFÁNTICA

¿Cómo traducir en una ecuación algebraica la determinación de los bits de la probabilidad de detención W de la que se habla en el articulo? El método utiliza una técnica desarrollada por Jones y Matijasevic, que a su vez se basa en el teorema de Lucas. Este afirma (fig. 1) que K «implica» N bit a bit si y sólo si el K-ésimo coeficiente binomial de orden N es impar. Jones y Matijasevic demuestran que esto equivale a decir que en la base b = 2N la K-ésima cifra de 11N es impar.

Matemáticamente esto se expresa así: K «implica» N si y sólo si existen enteros positivos o nulos únicos, b, x, y, z, u, v, w tales que:

$b = 2^N$

$(b + 1)^N = xb^{K+1} + y b^K + z$

$z + u + 1 = b^K$

$y + v + 1 = b$

$y = 2w + 1$

Para obtener una ecuación diofántica, basta reescribir estas cinco ecuaciones para que el miembro de la derecha sea 0, elevarlas al cuadrado y sumarlas. Se obtiene la ecuación

$$[b - 2^N]^2 = [(b + 1)^N - xb^{K+1} - y b^K - z]^2 + [z + u + 1 - b^K]^2 + [y + v + 1 - b]^2 + [y - 2w - 1]^2 = 0$$

La ecuación de 200 páginas que he obtenido se ha construido utlizado repetidamente esta técnica a fin de expresar en una ecuación diofántica el cálculo del N-ésimo bit de una K-ésima aproximación de W. Esta ecuación posee exactamente una solución si este bit es 1 y no posee ninguna si es 0. Se cambia entonces de punto de vista y se considera N no como un parámetro sino como una incógnita suplementaria. Para un determinado valor del parámetro N, la misma ecuación tendrá entonces un número finito o infinito de soluciones según que el N-ésimo bit de W sea 0 o 1 (el valor de K puede diferir de una solución a otra). Para K lo bastante grande, la aproximación de W es suficientemente buena para que el N-ésimo bit de la K-ésima aproximación de W sea el bueno. Pero es imposible calcular, para un determinado N, el valor de K a partir del cual el bit tiene dicho valor, pues la probabilidad de detención W es algorítmicamente aleatoria.

trado por el francés Edouard A. Lucas hace más de un siglo, que resuelve el 10º problema de Hilbert bastante fácilmente si se utiliza apropiadamente. El teorema de Lucas tiene que ver con la paridad de los coeficientes del binomio. Preguntémonos si el coeficiente de X^K en el desarrollo de $(1 + X)^N$ es par o impar, es decir, preguntémonos por la paridad del K-ésimo coeficiente binomial de orden N (para K = 0,1,2,...,N). El teorema de Lucas responde que este coeficiente es impar si y sólo si «*K implica N en tanto que sucesiones de bits*». Esto significa que este coeficiente es impar si a cada «1» de la representación binaria de K le corresponde un «1» en el mismo lugar en la representación binaria de N (fig. 1). De lo contrario, el coeficiente binomial es par. Utilizando la técnica de Jones y Matijasevic (véase el recuadro «Jugar a cara o cruz por medio de una ecuación diofántica»), basada en este notable teorema de Lucas, he creado un conjunto de programas escritos en los lenguajes C y SETL2* y los he puesto en acción en un ordenador IBM RIS System/6000. ¿Para obtener qué? Una ecuación diofántica, o más exactamente una ecuación diofántica exponencial. Las ecuaciones de esta clase sólo comprenden sumas, productos y exponenciaciones; las constantes y las incógnitas consideradas eran números enteros positivos o nulos. Contrariamente a una ecuación diofántica clásica, se admite que la potencia a la que se eleva una incógnita también puede ser una incógnita. Así, una tal ecuación puede contener no sólo términos como X^2 o X^3 sino también términos como X^Y o Y^X.

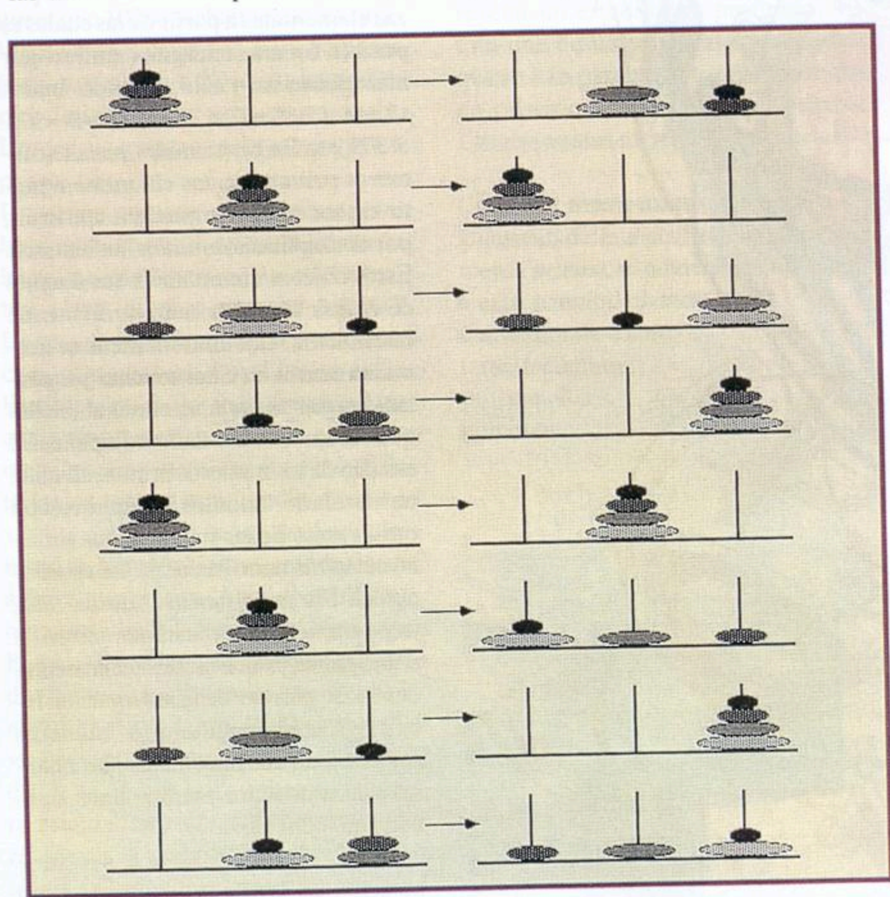
Más de doscientas páginas. La ecuación diofántica que he obtenido tiene más de 17 000 variables y ocupa más de 200 páginas.[II] Contiene un parámetro único, el número N. Para todo valor de este parámetro, hagámonos la pregunta siguiente: «*¿Tiene esa ecuación un número finito o infinito de soluciones en números enteros (es decir, un número finito o infinito de listas de 17 000 números enteros, siendo cada lista una solución de la ecuación)?*» La respuesta resulta ser un hecho aritmético aleatorio, análogo a tirar una moneda al aire. Es una transcripción aritmética del hecho matemático irreductible de que el N-ésimo bit de la probabilidad de detención Ω es 0 o 1: si esta ecuación diofántica (de parámetro N) tiene un número finito de soluciones, entonces este N-ésimo bit es 0; si la ecuación posee un número infinito de soluciones, este bit es 1 (subrayemos de paso que si no hay solución, el número de soluciones es finito y vale 0). En consecuencia, la respuesta a la pregunta no puede calcularse, y el N-ésimo bit de Ω tampoco. Esto no significa que los bits de Ω no estén definidos y determinados matemáticamente, sino más bien que no existe ningún algoritmo con un número finito de etapas para calcularlos, y que el conocimiento de los N primeros bits de Ω no ayuda en modo alguno a la determinación de los siguientes.

Los matemáticos están coincidiendo con sus colegas de la física teórica, lo cual tal vez no es mala cosa

La diferencia con el problema planteado por Hilbert es doble: por una parte, Hilbert sólo pensaba en las ecuaciones diofánticas clásicas, no exponenciales; por otra, la pregunta que había planteado era: «*¿Tiene solución la ecuación?*» Esta pregunta es indecidible, pero la respuesta no es totalmente aleatoria, sólo lo es en cierta medida. Las respuestas no son independientes entre sí; en efecto, es sabido que dado un número finito de ecuaciones diofánticas es posible determinar cuáles tienen solución si se sabe cuántas la tienen. Para obtener un azar realmente total, parecido al de un juego de cara o cruz, la pregunta adecuada es: «*¿Hay un número finito o infinito de soluciones?*». Mi aserción es que nunca podremos saberlo, pues decidir si el número de soluciones es finito o infinito, para cada valor de N, es un hecho matemático irreductible. La respuesta es algorítmicamente aleatoria. La única manera de progresar consiste en considerar las respuestas como axiomas. Si tratamos de resolver M veces la cuestión de saber si el número de soluciones es finito para M valores dados del parámetro N, habrá que incluir M bits de información en los axiomas de nuestro sistema formal. Es en este sentido preciso que

se puede decir que las matemáticas contienen «azar». En el sexto problema propuesto por Hilbert, la axiomatización de la física debía, según él, englobar la teoría de probabilidades. Con el tiempo, sin embargo, la teoría de probabilidades se ha convertido en una rama de las matemáticas de pleno derecho. Pero según lo que antecede, una forma extrema de «azar» —más exactamente de irreductibilidad— aparece en otro contexto, en matemáticas puras, en teoría elemental de números. Las investigaciones que llevan a estas conclusiones prolongan los trabajos de Gödel y Turing, quienes refutaron la hipótesis básica de Hilbert y otros según la cual toda cuestión matemática posee una respuesta unívoca.

El edificio de los números enteros.
Desde hace aproximadamente un siglo la filosofía y los fundamentos de las matemáticas despiertan un gran interés. Antes, se habían dedicado muchos esfuerzos a rigorizar el análisis matemático (las nociones de número real, de límite, etc.). El examen moderno de las matemáticas empezó realmente, creo yo, con la teoría del infinito de G. Cantor y las paradojas y sorpresas que engendró, así como con los esfuerzos de matemáticos como Peano, Russell y Whitehead para dotar a las matemáticas de fundamentos sólidos y rigurosos. Se habían puesto muchas esperanzas en la teoría de conjuntos. Se habían intentado definir rigurosamente los números enteros 0,1,2,3,..., en términos de conjuntos. Pero resultó que la noción de conjunto puede dar lugar a todo tipo de paradojas (Bertrand Rusell dio un ejemplo famoso: «*El conjunto de todos los conjuntos que no forman parte de sí mismos*»; ¿forma parte este conjunto de sí mismo?).
La teoría de conjuntos es una parte fascinante y vital de las matemáticas, pero me parece que ha habido un cierto desencanto en relación con ella y que se ha producido un regreso a los 0,1,2,3,.. intuitivos. Por desgracia, los trabajos que he mencionado, y en particular los míos propios, hacen que el edificio de los números enteros parezca menos firme de lo que se creía. Siempre he pensado, y probablemente la mayoría de los matemáticos lo creen también, en una especie de universo platónico en el que reina una «*realidad matemática*» independiente de la realidad física. Así, la cuestión de saber si una ecuación diofántica tiene un número finito o infinito de soluciones tiene muy poco sentido concreto, pero en mi fuero interno siempre he estado convencido de que aún cuando nosotros no pudiéramos responder, Dios sí podría hacerlo.
Con estos descubrimientos, los matemáticos, en cierto sentido, están coincidiendo con sus colegas de la física teórica. No necesariamente es mala cosa. En la física moderna, el azar y la imprevisibilidad desempeñan un papel fundamental; el reconocimiento y la caracterización de este hecho, que podría percibirse a priori como una limitación, constituyen un progreso. Creo que ocurrirá lo mismo en matemáticas puras.



La torre de Hanoi, este juego consiste en una pequeña colección de discos y tres palitos sobre los cuales se pueden colocar los discos. Empezando con todos los discos sobre el palito de la izquierda, hay que moverlos al palito derecho. No se permite mover más de un disco a la vez, ni poner un disco grande sobre uno más pequeño.

GREGORY J. CHAITIN trabaja en el centro de investigaciones Thomas J. Watson de IBM en Yorktown Heights, en Estados Unidos. Sus investigaciones versan sobre la teoría algorítmica de la información, de la que sentó las bases a mediados de los años 1960.

Mundo científico ha publicado:
(I) Hermann Haken, Arne Wunderlin, «El caos determinista», diciembre de 1990.
(II) Jean-Paul Delahaye, «Una extensión espectacular del teorema de Gödel: la ecuación de Chaitin», septiembre de 1988.

(1) G.J. Chaitin, *Advances in Applied Mathematics, 8*, 119, 1987; G.J. Chaitin, *Algorithmic Information Theory*, Cambridge University Press, 1990 (tercera impresión).
(2) J.P. Jones y Y.V. Matijasevic, *Journal of Symbolic Logic, 49*, 818, 1984.

*SETL2
es un lenguaje de programación que permite escribir estos programas de una manera más corta y fácil de comprender (aunque son más lentos). Este lenguaje se basa en una idea de J.T. Schwartz, del Instituto Courant de Nueva York, según la cual la teoría de conjuntos puede convertirse directamente en un lenguaje de programación (véase W.K.Snyder, *The SETL2 Programming Language*, Courant Institute, 1990; J.T. Schwartz et al., *Programming with Sets, An Introduction to SETL*, Springer-Verlag, 1986).

PARA MÁS INFORMACIÓN:

- E. Nagel, J.R. Newman, K. Gödel y J.-Y. Girard, *Le Théorème de Gödel, Seuil*, 1989.
- G.J. Chaitin, *Algorithmic Information Theory*, Cambridge University Press, 1990 (segunda reedición).
- G.J. Chaitin, *Information, Randomness and Incompleteness - Papers on Algorithmic Information theory*, World Scientific, 1990 (reedición).
- M.T. Hortalá González, J.L. Albert y M. Rodríguez Artalejo, *Matemática discreta y lógica matemática*, Ed. Complutense, Madrid, 1998.
- G.J. Chaitin, *The Unknowable*, Springer Verlag, Singapur, 1999.

La intriga de los números primos

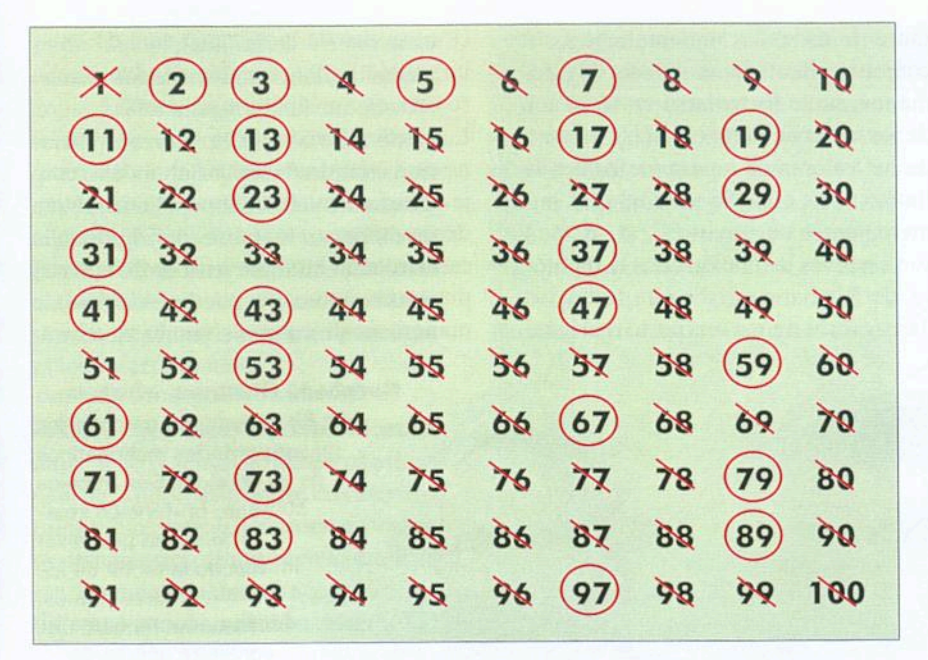
Henri Cohen

Los números primos, joyas que fascinan a los orfebres de la aritmética desde la Antigüedad, serán tal vez los últimos que entrarán en el paraíso del conocimiento matemático. La factorización y la hipótesis de Riemann son los dos grandes desafíos planteados a los profesionales de los números primos.



Los números primos se definen por una propiedad elemental: se dice que un entero es primo si sólo es divisible por 1 y por sí mismo. La sucesión de los números primos comienza por 2, 3, 5, 7, 11, 13, 17, 19, etc. (por convención, el número 1 no se considera primo). Se trata de piezas elementales a partir de las cuales es posible formar cualquier entero por multiplicación, y ello de modo único: $12 = 2^2 \times 3, 90 = 2 \times 3^2 \times 5, 49\,649 = 131 \times 379$, etc. En cierto modo, pues, los números primos son los «átomos» a partir de los cuales se pueden construir, por multiplicación, todos los enteros. Estos objetos matemáticos tan simples de definir han sido siempre fuente de fascinación. Los matemáticos se formulan acerca de ellos muchas preguntas, algunas todavía no resueltas, de las que veremos unos cuantos ejemplos. El estudio de los números primos, además, ha revelado vínculos profundos con otras ramas de las matemáticas, un fenómeno bastante frecuente en esta disciplina. Ha tenido incluso, desde 1980, importantes repercusiones sobre la criptografía, ya que ciertas técnicas modernas de cifrado de la información recurren esencialmente a los números primos. Una vez más, incluso las matemáticas más puras pueden tener aplicaciones concretas...

La criba de Eratóstenes. Para empezar, observemos la sucesión de los números primos, por ejemplo hasta 100. Esta lista se fabrica fácilmente (para números pequeños) por medio de una receta, conocida desde la Antigüedad,



1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figura 1. La criba de Eratóstenes es un método elemental para encontrar los números primos no demasiado grandes, por ejemplo, inferiores a 100. En la lista de todos los enteros inferiores a dicho límite se empieza por tachar los múltiplos de 2, que es el primer número primo. Luego se considera el primer número primo no tachado y se tachan sus múltiplos. Y así sucesivamente. Al término de la operación, los términos no tachados son primos.

llamada criba de Eratóstenes en honor del sabio griego del siglo III de nuestra era; en la lista de todos los enteros menores que 100 se tachan todos los múltiplos de 2 (es decir, todos los números pares no iguales a 2), luego los múltiplos de 3, los de 5 y así sucesivamente. Los únicos números sobrantes son los primos (fig. 1).

Una de las primeras cuestiones que cabe plantearse en relación con la lista de los números primos es la de su longitud. ¿Hay infinitos números primos? La respuesta es que sí. La demostración, contenida en *Los Elementos* de Euclides, es de lo más simple. Para construir un número primo mayor que un entero n se calcula el producto P de todos los enteros comprendidos entre 1 y n y se añade 1 al resultado. Se obtiene así un número N (muy grande) igual a P + 1 que no puede ser divisible por ningún número inferior a *n* (en efecto, P, por construcción, es divisible por todos los números menores que *n*, pero 1 no lo es). Por lo tanto, *N* es primo o divisible por un número primo mayor que n; en ambos casos existe un número primo mayor que n y ello cualquiera que sea el valor de *n*. De ahí la existencia de infinitos números primos.

Salta a la vista una segunda propiedad al examinar la lista de los números primos: lo irregular de su distribución. Por ejemplo, no hay ningún número primo entre 114 y 126, pero hay cinco entre 97 y 109. Estos números no parecen regidos por ningún orden. ¿Es realmente así, o existe un orden sutil que podría revelar un estudio más profundo? Es ésta una de las principales cuestiones que se han planteado los matemáticos en relación con los números primos. Las respuestas todavía son parciales.

Cada vez más escasos. Pese a la irregularidad de la distribución de los números primos, se advierte en ellos una cierta propiedad media. Se puede constatar fácilmente que los números primos son cada vez más raros. Por ejemplo, hay 168 números primos entre 0 y 1 000, 106 entre 10 000 y 11 000, 81 entre 100 000 y 101 000, etc. y sólo dos entre 10^{100} y $10^{100} + 1\,000$.

Valor aproximado. Como veremos, la cada vez mayor escasez de los números primos es una propiedad bastante bien comprendida en la actualidad. Tradicionalmente se representa por $\pi(x)$ (que nada tiene que ver con π = 3,14159..) el número de números primos menores o iguales que x. En 1798, examinando extensas tablas obtenidas por él mismo y por otros, el alemán Carl Friedrich Gauss había formulado la hipótesis de que para valores grandes de x, la función $\pi(x)$ es aproximadamente igual a x, donde $\ln x$ designa el logaritmo natural de x (esto es, $y = \ln x$ significa que $e^y=x$, donde e = 2,71828...). Un cálculo por ordenador permite comprobar, en efecto, la proximidad de las funciones $\pi(x)$ y $x/\ln x$ (fig. 2).

Hacia 1859, el alemán Bernhard Riemann introdujo la célebre función zeta, directamente relacionada con los números primos

En el siglo XIX, muchos matemáticos trataron de demostrar la conjetura de Gauss. El matemático ruso Pafnuti Chebichev, por ejemplo, demostró por métodos elementales que existen dos constantes positivas $c_1 < 1$ y $c_2 >$ 1 tales que:

$$c_1\, x/\ln x < \pi(x) < c_2\, x/\ln x.$$

Pero fue el matemático alemán Bernhard Riemann quien, hacia 1859, realizó el descubrimiento más espectacular. La idea de Riemann con-

sistía en considerar la función $\zeta(s) = 1 + 1/2^s + 1/3^s + 1/4^s + ...$ Esta célebre función zeta de Riemann está directamente relacionada con los números primos; mediante simples manipulaciones algebraicas se demuestra, en efecto, que se cumple $1/\zeta(s) = (1 - 1/2^s)(1 - 1/3^s)(1 - 1/5^s) ...$, un producto en el que aparecen sólo los números primos (2, 3, 5, 7, 11, etc.).

tante de todas las matemáticas. Esta conjetura, llamada «hipótesis de Riemann», suele formularse en términos de los ceros de la función $\zeta(s)$, esto es, de los valores de s para los cuales la función ζ es nula. Se sabe que los números pares negativos (-2, -4, -6, etc.) son ceros de la función zeta: la hipótesis de Riemann afirma que todos los demás ceros tienen una parte real igual

ción vendrá de la física, lo cual no es imposible. Pero también es posible, claro está, que la hipótesis sea falsa.

La distribución de los números primos no es el único tema que abunda en conjeturas. Es sabido el papel crucial que desempeñan los números primos en la estructura multiplicativa de los enteros, pues todo entero puede escribirse de manera única como producto de nú-

El alemán Carl Friedrich Gauss fue matemático, físico y astrónomo. En sus múltiples trabajos de matemáticas figuró también el estudio de los números primos. Este científico formuló la hipótesis de que el número de números primos inferiores a un valor grande x es aproximadamente igual a x/ln x, donde ln x es el logaritmo neperiano (o natural) de x. la conjetura de Gauss fue demostrada cuarenta años después de su muerte: es el famoso «teorema de los números primos».(Foto BPK)



Bernhard Riemann, alumno de Gauss, figura entre los más grandes matemáticos de su siglo. Sus trabajos sobre la «función zeta» hicieron progresar mucho la teoría de los números primos. Riemann formuló una conjetura acerca de esta función conocida como «hipótesis de Riemann», que tiene profundas y múltiples repercusiones en varias ramas de las matemáticas pero que todavía no ha podido demostrarse. (Foto BPK)

La intuición. La función zeta ya había sido considerada por Bernoulli y Euler, pero sólo para valores reales de s. La gran intuición de Riemann fue que el conocimiento de la función $\zeta(s)$ generalizada a valores complejos de s permitiría obtener informaciones más finas sobre los números primos, como así fue, efectivamente. Dicho en términos más precisos, Riemann demostró que la función $\pi(x)$ está íntimamente ligada a la función z para valores complejos de s e incluso encontró que una mejor aproximación a $\pi(x)$ es la función logaritmo integral $Li(x)$ (fig. 2), lo que confirmaba otra predicción de Gauss. Pero hubo que esperar a 1896 para que el francés Jacques Hadamard y el belga Charles de la Vallée-Poussin, independientemente el uno del otro, colmaran algunas lagunas de la argumentación de Riemann y demostraran la conjetura de Gauss, conocida desde entonces como «teorema de los números primos»: cuando x tiende a infinito, el cociente entre $\pi(x)$ y $x/\ln x$ tiende a 1.

El método de Riemann, completado por sus sucesores hasta nuestros días, ha abierto más incógnitas de las que ha resuelto. En el marco de estas investigaciones, Riemann había enunciado la conjetura no demostrada más impor-

a 1/2, es decir, que tienen la forma: $1/2 + iy$, donde y es un número real e $i = \sqrt{-1}$. Esta formulación tiene el inconveniente de no hacer mención de los números primos. Pero dado que se puede pasar de enunciados relativos a ζ a enunciados relativos a los números primos, hay otra formulación de la hipótesis de Riemann que sí menciona explícitamente la distribución de los números primos: la función $\pi(x) - Li(x)$, que mide la distancia entre el valor exacto de $\pi(x)$ y la predicción de Gauss, no crece más deprisa que a>1/2 para todo número a > 1/2. En otras palabras, la hipótesis de Riemann equivale a afirmar que para todo $a > 1/2$, el límite cuando x tiende a infinito de la expresión $\pi(x) - Li(x))/x^a$ es igual a cero.

Aunque tal cosa no sea evidente aquí, la hipótesis de Riemann es absolutamente fundamental. Su demostración tendría profundas repercusiones no sólo en aritmética sino también en otros campos de la matemática y también en física teórica. Un gran número de enunciados matemáticos dependen de ella. Y pese a los notables resultados obtenidos en relación con esta hipótesis, nadie sabe de qué lado vendrá el ataque decisivo. Algunos físicos teóricos piensan incluso que la idea de la demostra-

meros primos. Pero ¿qué ocurre cuando se suman números primos? La primera conjetura en esa dirección fue enunciada por el matemático alemán Christian Goldbach en una carta dirigida en 1742 a Leonhard Euler. La conjetura de Goldbach, no demostrada todavía, dice que todo número par mayor o igual que cuatro es igual a la suma de dos números primos (por ejemplo 16 = 11 + 5). Este enunciado fue rápidamente generalizado a los impares: todo número impar mayor o igual que 3 sería la suma de un número máximo de tres números primos.

Según la conjetura de Goldbach, todo número par mayor o igual que cuatro es la suma de dos números primos

No más de siete. A diferencia de la hipótesis de Riemann, la conjetura de Goldbach no tiene ningún interés intrínseco, exceptuando, claro está, el reto que plantea. Es probable, por otra parte, que estemos mucho más cerca de demostrar la conjetura de Goldbach que la de Riemann. Por ejemplo, el so-

viético Iván Matvéievich Vinogradov demostró hacia 1937 que todo número impar *lo bastante grande* es la suma de un máximo de tres números primos y que todo número par *lo bastante grande* es la suma de un máximo de cuatro números primos. Por otra parte, Olivier Ramaré demostró en 1991 que todo entero es la suma de un máximo de siete números primos (no más de seis para los enteros pares).[2]

Infinitos gemelos. Otro problema similar al de la suma de números primos es el de la diferencia entre dos números primos sucesivos. El examen de una lista de números primos revela unos pares de números, llamados gemelos, que difieren en dos unidades, como (5,7), (29,31) y (281,283). La conjetura de los números primos afirma que hay infinitos. Se cree incluso que el número de números primos gemelos inferiores a x es del orden de $c_3 x/(\ln x)^2$, para una constante conocida c_3.

También es posible interesarse por la mayor distancia posible entre dos números primos sucesivos. La situación es aquí infinitamente menos clara. En 1937, el matemático sueco Harald Crámer formuló la hipótesis de que siempre existe un número primo entre x y $(\ln x)^2$ (o tal vez entre x y $x + (\ln x)^a$ para todo $a > 2$).[3]

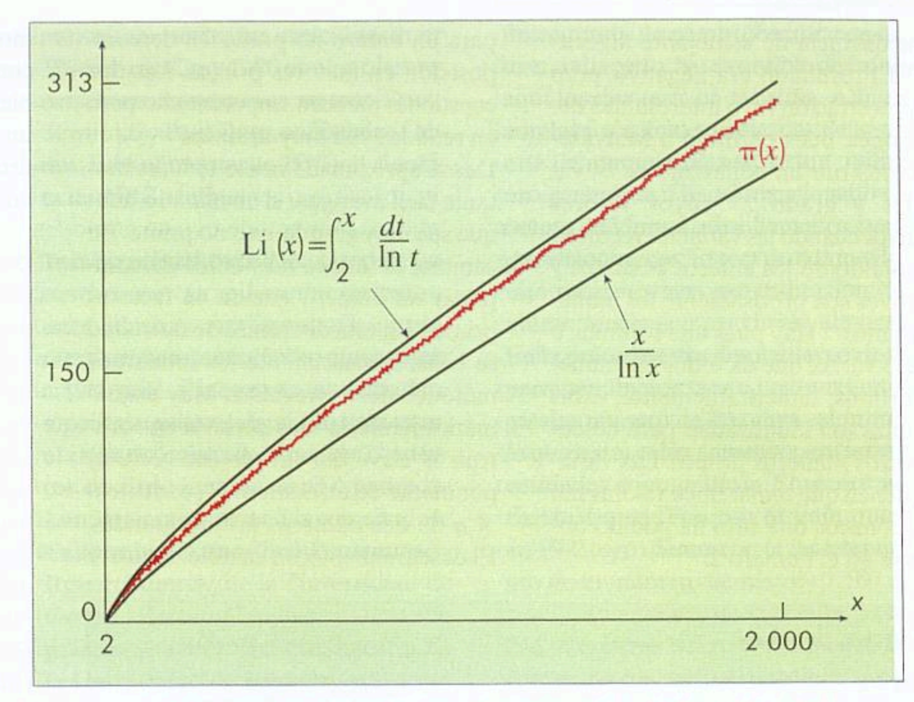


Figura 2. Esta gráfica muestra un fragmento de la función π(x), que representa el número de números primos inferiores a *x*. Esa función es muy irregular, a imagen y semejanza de los propios números primos. No obstante, cuando *x* se hace grande, *x*/ln*x* se comporta en promedio como *x*/ln*x* y el logaritmo integral Li(*x*).

Estamos, sin embargo, muy lejos de poder demostrar resultados de este tipo, ni siquiera suponiendo verdaderas otras conjeturas como la de Riemann. Un importante resultado lo obtuvieron en 1994 los británicos Roger Baker y Glyn Harman.[4] Afirma el resultado que para todo x lo bastante grande existe por lo menos un número primo entre x y $x + x^{0.535}$. La hipótesis de Riemann daría x y $x + x^{1/2}$, un resultado mucho más fino pero todavía alejado de la conjetura de Cramer.

Esta investigación está en la línea de otro viejo sueño de los matemáticos, la obtención de una fórmula que proporcione todos los números primos. Un primer paso en esa dirección fue dado en 1826 por el alemán Peter Gustav Lejeune-Dirichlet, quien demostró que si a y b son números sin otros factores comunes que el 1 (números primos entre sí), existen infinitos números primos de la forma $an + b$, donde n es un entero.

Propiedades elementales. Cabría la tentación de generalizar tales resultados a polinomios en n de grado superior a 1; de hecho no se sabe en absoluto cómo hacerlo. Por ejemplo, hay una conjetura que afirma la existencia de infinitos números primos de la forma $n^2 + 1$; pero nadie sabe cómo demostrarla. Esto no ha sido óbice para que los matemáticos hayan descubierto algunos polinomios sorprendentes, como $n^2 - n + 41$; cuando n está comprendido entre –40 y 40 (ambos incluidos), este valor tiene como valor un número primo (pero se ignora si hay infinitos números primos de esta clase). Esta propiedad es algo más que un divertimento numérico, pues al estar relacionada con la teoría de los «cuerpos de números algebraicos» se la comprende perfectamente bien.[5] Hay varios ejemplos del mismo tipo.

Pero se han logrado cosas todavía más espectaculares. En 1900, el matemático alemán David Hilbert había propuesto en el Congreso internacional de matemáticos 23 problemas sin resolver (véase el artículo de J.-P. Boruguignon e H. Sinaceur en este número) . El décimo consistía en encontrar un algoritmo universal que permitiera averiguar si una ecuación algebraica tiene o no soluciones enteras. Como conclusión de numerosos trabajos, el matemático ruso Yuri Matyasevich demostró en 1969

la inexistencia de semejante algoritmo.[6] El resultado era negativo, pero su demostración tuvo interesantes derivaciones, pues permitió a Matyasevich construir un polinomio de 24 variables y grado 27 cuyos valores positivos cuando las variables recorren el conjunto de los enteros positivos y negativos son precisamente los números primos.[7] Hay pues una fórmula, o incluso varias, que da todos los números primos; lamentablemente, estas fórmulas son inutilizables para descubrir otros números primos. Las variables tienen que tomar unos valores tan astronómicos que sólo ha podido obtenerse así el número 2.

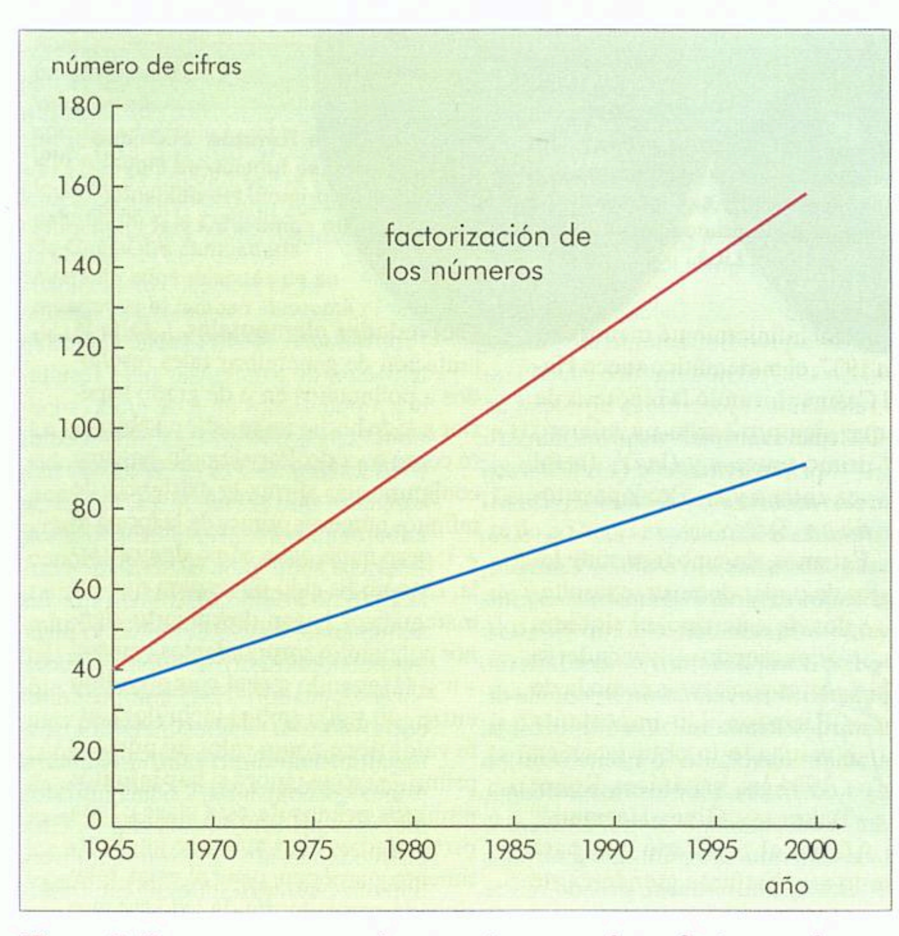


Figura 3. Descomponer un número entero grande en factores primos es mucho más difícil que encontrar un nuevo número primo. Esta gráfica muestra la evolución del tamaño de los números que se han conseguido factorizar (medido en número de cifras).
En azul: lo que se es capaz de hacer en un tiempo razonable, es decir, con un ordenador de tipo estación de trabajo utilizado durante varios días como máximo.
En rojo: lo que se es capaz de hacer un tiempo no razonable, es decir, haciendo trabajar durante meses una red de miles de ordenadores combinados con algunos superordenadores.

Ahora bien, el hallazgo de nuevos números está estrechamente ligado a dos problemas en torno a los cuales se despliega hoy en día una actividad bastante intensa. Uno de ellos consiste en averiguar, para un entero positivo, si es o no primo. Otro consiste en determinar, para un entero no primo, su descomposición en factores primos. Las dos cuestiones parecen emparentadas pero en realidad son muy distintas.

Desde aproximadamente 1980, es bastante fácil averiguar si un número, aunque sea muy grande, es o no primo. En cambio, es mucho más difícil descomponer un número grande en factores primos. En esta diferencia de dificultad se basan precisamente los modernos métodos de criptografía. Muy esquemáticamente, la idea consiste en construir la clave del cifrado (no secreta) por medio de un número N producto $p \times q$ de dos números primos grandes. El desciframiento, en cambio, requiere el conocimiento de p y de q por separado, unos valores que sólo se comunican a las personas autorizadas. Para descifrar un mensaje en clave, un espía tendría que encontrar, conociendo N, sus dos factores primos p y q. Ahora bien, para N lo bastante grande (digamos que del orden de 150 cifras) y convenientemente elegido, es imposible realizar esta operación inversa en un tiempo aceptable (véase el recuadro «Criptografía de clave pública y números primos»).

De Eratóstenes a Fermat. Consideremos la primera cuestión mencionada: dado un entero n mayor que 1, determinar si es o no primo. Una primera aproximación consiste en imitar la criba de Eratóstenes y tratar de dividir n por todos los números primos inferiores. Si uno de ellos divide a n entonces n no es primo; en caso contrario, sí lo es. Aunque fastidioso para valores grandes de n, el método queda facilitado por las dos consideraciones siguientes. En primer lugar, es inútil tratar de dividir por todos los números primos inferiores a n. Es posible detenerse en $\sqrt{n}$ porque si uno de los factores es mayor que $\sqrt{n}$ el 14 complementario será necesariamente menor. En segundo lugar, no siempre se dispone de una tabla de números primos lo bastante extensa. En tal caso, basta tratar de dividir por 2 y por todos los números impares menores que $\sqrt{n}$. Estos, por supuesto, son mucho más numerosos que los números primos, pero se evita así tener que recurrir a una tabla de números primos.

En criptografía se utiliza a veces una de las dificultades de la disciplina: descomponer un número lo bastante grande en factores primos

En cualquier caso, este método ingenuo para comprobar el carácter primo de un número n funciona razonablemente bien hasta un n del orden de 10000 y con un ordenador hasta $n = 10^{14}$. Tales valores son claramente insuficientes, especialmente para la criptografía. ¿Qué más se puede hacer?

El segundo enfoque, mucho más eficaz, es debido a Pierre de Fermat, el célebre matemático francés de la primera mitad del siglo XVII. Se basa en el «segundo teorema de Fermat», cuyo enunciado es como sigue: si n es un número primo y a un entero no divisible por n, entonces $a^{n-1} - 1$ es divisible por n.

El interés de este resultado es doble, En primer lugar, comprobar si $a^{n-1} - 1$

es divisible por n no es difícil, ni siquiera cuando n es un número enorme de cientos de cifras. Por otra parte, aunque el teorema no prohíbe la existencia de números n no primos que satisfagan la misma condición, tales números son raros.
Por tanto, un método razonable para comprobar el carácter primo de un número impar n es el siguiente: se comienza dividiendo n por los números impares 3, 5, 7, hasta un límite bastante pequeño. Si n no es divisible por ninguno de ellos, se sospecha que es primo y se examina la divisibilidad de $a^{n-1} - 1$ por n (como hemos dicho, esto se puede hacer rápidamente). Si $2^{n-1} - 1$ no es divisible por n, el teorema de Fermat implica inmediatamente que n no es primo. Si $2^{n-1} - 1$ es divisible por n, no se puede llegar todavía a ninguna conclusión, aunque es bastante probable que en tal caso n sea primo.

Se han creado algoritmos para averiguar el carácter primo de números cualesquiera con cientos de cifras

El método anterior, basado en el resultado de Fermat, constituye un progreso esencial, no obstante lo cual presenta un inconveniente importante. Permite demostrar eventualmente que un número grande no es primo, pero no afirmar que es primo un número n tal que $2^{n-1} - 1$ sea divisible por n. ¿Cómo demostrar que un número grande es efectivamente primo, cuando tal es el caso? Desde el siglo XVII se han realizado numerosos progresos en esta dirección. Por ejemplo, en 1878, el matemático francés Edouard Lucas halló un test parecido al de Fermat pero válido para unos números muy especiales, los números de Mersenne (números de la forma $2_q - 1$, donde q es un entero positivo). El test de Lucas, mejorado en 1930 por el norteamericano Derrick H. Lehmer, es muy fácil de poner en práctica, especialmente con ordenador, y permite demostrar el carácter primo de números gigantescos. Los mayores números primos conocidos son casi todos números de Mersenne. El test de Lucas-Lehmer, por lo demás, se utiliza sistemáticamente para comprobar la fiabilidad de los superordenadores Cray. El mayor número primo actualmente conocido fue descubierto en 1994 con un Cray C-90 por los célebres cazadores de números primos David Slowinski y Paul Gage, en Estados Unidos. Este número vale $2^{859\,433} - 1$ y consta de 258 716 cifras.

Test de primalidad. Pero el test de Lucas-Lehmer se aplica únicamente a unos números muy especiales, los números de Mersenne. Hubo que esperar a 1980 para disponer de un verdadero algoritmo general, capaz de demostrar el carácter primo de números de cientos de cifras. La idea inicial partió de Len Adelman, de la Universidad de California del Sur, Carl Pomerance y Robert Rumely, de la Universidad de Georgia (Estados Unidos), y la versión práctica se debe a Hendrik Lenstra, de la Universidad de Berkeley y al autor de estas líneas.[8] El test se conoce como test APRCL, por las iniciales de los nombres de sus inventores.
En 1980, Oliver Atkin, de la Universidad de Illinois (Estados Unidos) y François Morain, de la Escuela Politécnica de Palaiseau (Francia), crearon un algoritmo muy distinto de similar eficacia.[9] Su ventaja estriba en que cuando declara primo a un número la afirmación puede comprobarse fácilmente mediante los instrumentos de cálculo que suministra. En el caso del test APRCL, en cambio, el único modo de comprobación consiste en rehacerlo. No es posible describir aquí los dos tests en cuestión. Ambos emplean técnicas sofisticadas de la teoría de números. El test APRCL utiliza una generalización del mencionado teorema de Fermat a los llamados cuerpos ciclotómicos. El test de Atkin-Morain recurre a una generalización del mismo teorema a las curvas elípticas.
En la práctica, estos dos modernos tests permiten demostrar en pocos minutos que un número de 200 cifras es primo, lo que basta sobradamente para las necesidades criptográficas. El récord actual lo detenta François Morain, quien en 1992 logró demostrar el carácter primo de un número de más de 1 500 cifras.[10] Aunque se está muy lejos de las 258 716 cifras del mayor número de Mersenne conocido, el resultado es notable, pues este número no tienen ninguna propiedad especial.

Problema inverso. Así pues, se puede considerar que el problema de demostrar que un número es primo está zanjado satisfactoriamente desde 1980.



Criptografía de clave pública y múmeros primos

¿Cómo transmitir mensajes secretos asegurándose de que no serán comprendidos por un eventual enemigo? Tal es el objetivo de la criptografía. En principio, el mensaje a transmitir tiene que codificarse previamente por medio de una clave que el emisor y el receptor mantienen en secreto. En general, esta clave puede ser fácilmente «invertida», por lo que sirve tanto para cifrar el mensaje inicial como para descifrar el mensaje cifrado. En tal caso, que es el clásico, el emisor y el receptor comparten un mismo secreto, la clave que sirve para cifrar y descifrar. El principio de la criptografía de clave pública, inventado en 1976 por Whitfield Diffie y Martin Hellmann, de la Universidad de Stanford, es muy distinto.[12] El método supone que la clave del cifrado no puede ser fácilmente invertida para hallar la clave del desciframiento. La primera puede ser pública, mientras que sólo el receptor puede conocer la segunda. La seguridad es entonces mucho mayor.

Supongamos que Alicia quiere enviar a Bernardo un mensaje secreto por medio de una clave pública. En tal caso, Bernardo tiene que comunicar a Alicia una clave de cifrado que puede ser conocida por todo el mundo (es la clave pública). Alicia la utiliza para cifrar su mensaje, que luego envía a Bernardo. Para descifrar el mensaje en clave, Bernardo utiliza su clave privada, que es el único en conocer.

El sistema de cifrado de clave pública más conocido y utilizado es probablemente el sistema RSA, inventado en 1978 por Ronald Rivest, Adi Shamir y Leonard Adelman, del MIT (*Massachusetts Institute of Technology, Estados Unidos*).[13]

El método se basa en las dos constataciones siguientes:

1) Es relativamente fácil encontrar dos números primos grandes y multiplicarlos para dar un número $N = p \times q$ que sirva para una clave de cifrado.

2) El proceso inverso, esto es, la determinación de p y q a partir de un número grande N es imposible de realizar en un tiempo aceptable. Esta imposibilidad garantiza que un eventual espía no podrá descifrar el mensaje aunque conozca el número N.

¿Cómo funciona más exactamente el sistema RSA? He aquí un ejemplo ligeramente simplificado. Para que Alicia pueda enviar un mensaje a Bernardo con toda seguridad, Bernardo elige dos números primos grandes p y q cuya división por 3 dé como resto 2 (hay otras posibilidades además de estos valores 3 y 2), calcula $N = p \times q$ y transmite a Alicia el valor de N. Para enviar su mensaje, Alicia lo transforma en una sucesión de cifras y luego lo divide en segmentos de longitud aproximadamente igual a N.

Alicia envía cada fragmento por separado del modo siguiente. Supongamos que un fragmento es un número x: Alicia calcula el resto y de la división de x^3 por N y el segmento cifrado es simplemente el número y. Para descifrar el segmento recibido, Bernardo utiliza el número ultrasecreto $e = (2(p - 1)(q - 1) + 1/3$, el cual, debido a la elección de p y q, es entero. Se puede demostrar, efectivamente, que si y es un fragmento codificado siguiendo el método anterior, el fragmento original x es igual al resto de la división de y^e por N. El cálculo puede realizarse muy rápidamente. Como se ve, para descifrar el mensaje en clave ni siquiera es necesario acordarse de p y q, basta acordarse de e. Es muy importante que este entero e no caiga en poder de un eventual espía, pero como Bernardo es el único que lo conoce, los riesgos de indiscreción son muy reducidos.

La importancia de la teoría de números en criptografía es tal que los matemáticos se ven confrontados a problemas deontológicos. Por ejemplo, si uno de ellos descubre un método de factorización numérica mucho más eficaz que los precedentes, ¿qué debe hacer? ¿Comunicarlo al primer ministro, exponerlo públicamente en una conferencia internacional para que nadie pueda aprovecharse de él a expensas de otros o venderlo al mejor postor? Afortunadamente, hasta dónde se sabe, las mejoras encontradas por los matemáticos no son lo bastante revolucionarias como para que el problema se plantee en toda su acuidad.

Consideremos ahora el problema inverso. Supongamos que sabemos que n no es primo, por ejemplo porque $2^{n-1} - 1$ no es divisible por n. ¿Cómo encontrar los divisores de n? Digamos enseguida que este problema de la factorización, mucho más difícil, carece todavía de solución satisfactoria pese a los progresos realizados.

Es posible utilizar la misma estrategia ingenua de antes, que consistía en dividir el número por los números primos (o impares) hasta $\sqrt{n}$. Como hemos dicho, este método -que es el único que generalmente conocen los aficionados- puede aplicarse razonablemente con ordenador a números de hasta quince cifras. Más allá, el método es demasiado laborioso.

La factorización y la hipótesis de Riemann son los dos grandes desafíos planteados a los profesionales de los números primos. Pese a los numerosos progresos realizados, es imposible prever hoy en día cómo y cuándo se dará con la solución.

HENRI COHEN es profesor de la Universidad Bordeaux I e investiga en el Laboratorio de algorítmica aritmética y experimental. Se interesa por las aplicaciones de la informática a la teoría de números.

(1) I.M. Vinogradov, *Dokl. Akad. Nauk. SSR, 15*, 169, 1937.
(2) O. Ramaré, *Ann. Scuola Norm. di Pisa*, 1995.
(3) H. Cramér, *Acta Arith., 23*, 1937.
(4) R. Baker y G. Harman, *Proc. London Math. Soc.*, 1995.
(5) P. Ribenboim, *Ens. Math. 34*, 23, 1988.
(6) Y.V. Matyasevitch, *Dikl. Akad. Nauk. SSSR, 191*, 279, 1970.
(7) Y.V. Matyasevitch, *Dokl. Akad. Nauk. SSSR, 196*, 770, 1971; P.J. Jones *et al., Am. Math. Monthly, 83*, 449, 1976.
(8) H. Cohen y H.W. Lenstra, *Math. Comp., 42*, 297, 1984.
(9) O. Atkin y F. Morain, *Math. Comp., 61*, 29, 1993.
(10) F. Morain, *Rapport de recherche du laboratoire d'informatique de l'Ecole polytechnique (Palaiseau)*, 1992.
(11) A.K. Lenstra y H.W. Lenstra (eds.), *The Development of the Number field sieve*, Springer, 1993.
(12) D. Diffie y M. Hellmann, *IEEE Transactions on Information Theory, IT-22*, 644, 1976.
(13) R.L. Rivest *et al., Communications of the ACM, 21*, 120, 1978.

Mundo Científico ha publicado:
(I) H. Cohen y D. Nordon, «La aritmética asistida por la geometría y el ordenador», mayo de 1989.

Para más información:

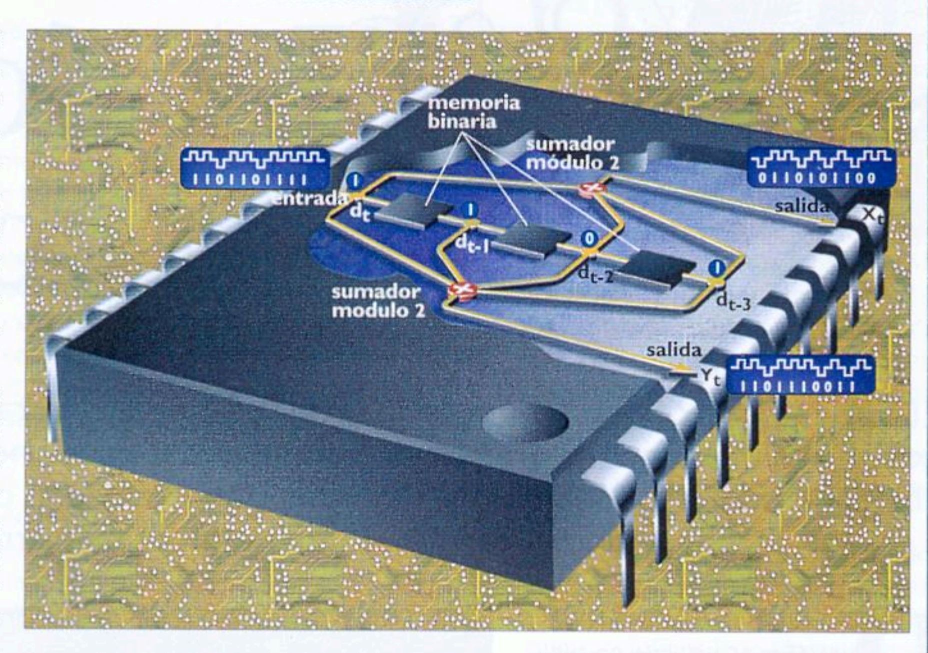
W.J. Ellison y M. Mendès-France, *Les nombres premiers*, Hermann, 1975.
H. Riesel, *Prime Numbers and Computer Methods for Factorization, Birkhäuser*, 1985.
E. Trillas y J. Gutiérrez Ríos, *Aplicaciones de la lógica borrosa*, Consejo Superior de Investigaciones Científicas, col. Nuevas Tendencias, Madrid, 1992.

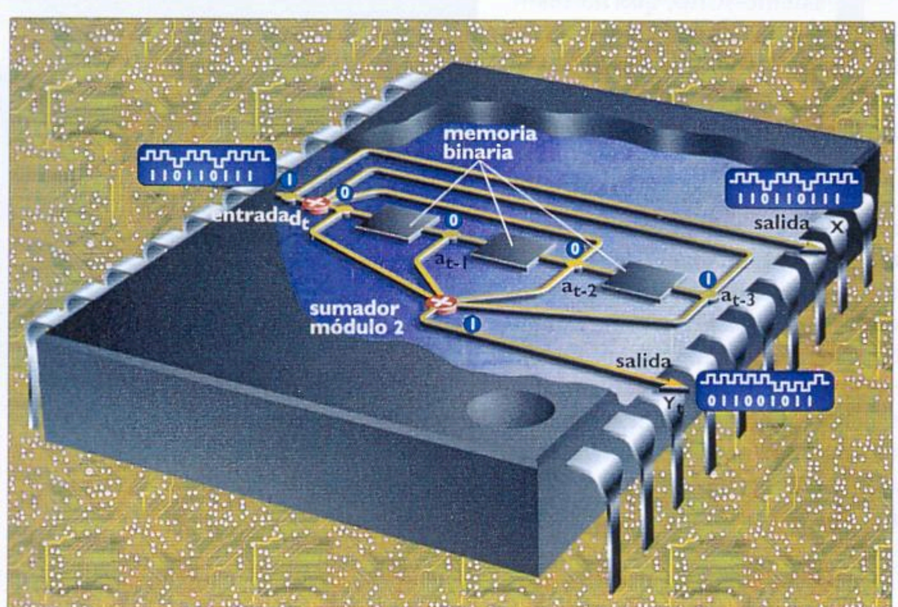
Desmultiplicar el Mensaje

¿Cómo introducir redundancia en un mensaje binario por medio de un cifrado convolutivo? El circuito de un codificador convolutivo está constituido por un registro de desplazamiento* que contiene, por ejemplo, tres memorias y dos sumadores módulo 2*. En todo momento, el codificador dispone del dato en el tiempo presente y de tres datos precedentes (dt-1, dt-2, dt-3).

Los dos sumadores representados aquí por ≈ no tienen en cuenta los mismos datos, por lo que dan otras dos secuencias completamente distintas de la primera. En cada operación, dan un «1» si el número de sus valores de entrada en «1» es impar y un «0» si dicho número es par. Por ejemplo, la secuencia 1101101111, cifrada teniendo en cuenta los datos primero, tercero y cuarto para la suma, dará la serie 0110101100. Si se tienen en cuenta los cuatro datos presentes en el registro en el instante t se obtiene la serie 1101110011. Son sinónimos de la secuencia original (A).

Por supuesto, es posible hallar una combinación de operadores lógicos que recupere la secuencia de partida. Así, se puede demostrar que la relación $X_t + X_{t-1} + X_{t-2} + X_{t-3} = Y_t + Y_{t-2} + Y_{t-3}$ (módulo 2) se verifica en todo momento. El descodificador se basará en esta propiedad algebraica para estimar el mensaje original {d}.





Cuando los errores de transmisión no son muy numerosos, se corrigen fácilmente con este cifrado. En cambio, si se producen paquetes de errores, el descodificador puede tener muchas dificultades para estimar correctamente ciertos datos. Naturalmente, el poder de corrección del código es tanto más elevado cuanto mayor es el tamaño del registro, pero la complejidad del descodificador asociado crece exponencialmente con el tamaño.

En el primer ejemplo, la secuencia inicial {d} no es transmitida directamente. Un código de este tipo se llama «no sistemático». Una versión «sistemática» del mismo código posee salidas modificadas (B).

boratories, esta teoría no fue la única en abrir lo que hoy se llama la era informática. Al mismo tiempo y en la misma empresa, Wiliam Shockley y su equipo experimentaron el primer transistor. Este doble *big bang*, una prodigiosa coincidencia, hizo nacer casi como gemelos el componente semiconductor —que según su estado de conducción, abierto o cerrado, es capaz de representar materialmente la información binaria 0 o 1— y el *shannon o bit* (contracción de *binary unit*), unidad de medida de dicha información. De entrada, con la miniaturización de los circuito lógicos que permitía el transistor, la teoría encontraba ya un vasto campo de aplicaciones.

Umbral de ruido. Uno de los resultados presentados por Shannon es particularmente sorprendente: en una transmisión numérica en presencia de perturbaciones, si el nivel medio de éstas no sobrepasa un cierto umbral, el receptor puede identificar el mensaje de origen sin error alguno.[1]

Para ello, basta codificar la información apropiadamente. Ilustremos el caso con un ejemplo sencillo. Dos personas se hablan cerca de una carretera donde la circulación es bastante intensa. El ruido estorba la conversación y los picos de las perturbaciones sonoras corresponden al paso de los vehículos. Supongamos, en un primer

tiempo, que una de las personas emite regularmente una letra —«a», «b»,...— elegida al azar entre las 28 letras del alfabeto. La probabilidad de obtener una determinada letra es igual para todas ellas: 1/28. Como no existe ninguna relación entre los sonidos emitidos, el oyente, si no lee en los labios, no podrá reconocer todas las letras con certeza y la transmisión se

En una transmisión, la detección de errores es posible si y sólo si hay un efecto de redundancia

verá salpicada de errores. Si, en un segundo tiempo, el hablante enuncia frases completas, conformes a la sintaxis de la lengua del oyente e integradas en un mensaje cargado de sentido, el ruido del tráfico entorpecerá mucho menos la transmisión. En efecto, el lenguaje contiene lo que se llama «redundancia». Las letras están integradas en palabras, las palabras en frases y las frases en un discurso. Generalmente la mayor parte del sentido transmitido puede reconocerse sin utilizar la totalidad de las frases: será suficiente un número limitado de palabras; las otras acompañan, confirman o matizan. Ahí reside la redundancia. Y esto es lo que Shannon anunciaba: en presencia de perturbaciones tiene que existir un sistema de codificación que permita anular completamente los efectos de las distorsiones y reconocer así el contenido del mensaje emitido, corrigiendo los errores.

Este resultado teórico ha revelado su importancia permitiendo la explosión del mercado de telecomunicaciones. En este campo, la mejora del poder de corrección de un código incide directamente sobre el coste de un sistema de transmisión. Esto significa una cobertura más amplia y un mayor alcance. En resumen: unas condiciones más severas que no afectan la calidad de la información recibida. Gracias a los avances actuales, el índice de error en los teléfonos móviles numéricos es del 1 por 10 000. Esta eficacia permite reducir el tamaño de las antenas, las potencias de emisión —y por tanto, el peso de las baterías de alimentación— e incluso el volumen de los paneles solares cuando se trata de sondas espaciales. Eficaces hasta los confines del sistema solar, estas últimas se benefician de un ahorro en su construcción y en su lanzamiento, que puede cifrarse en decenas de millones de dólares.

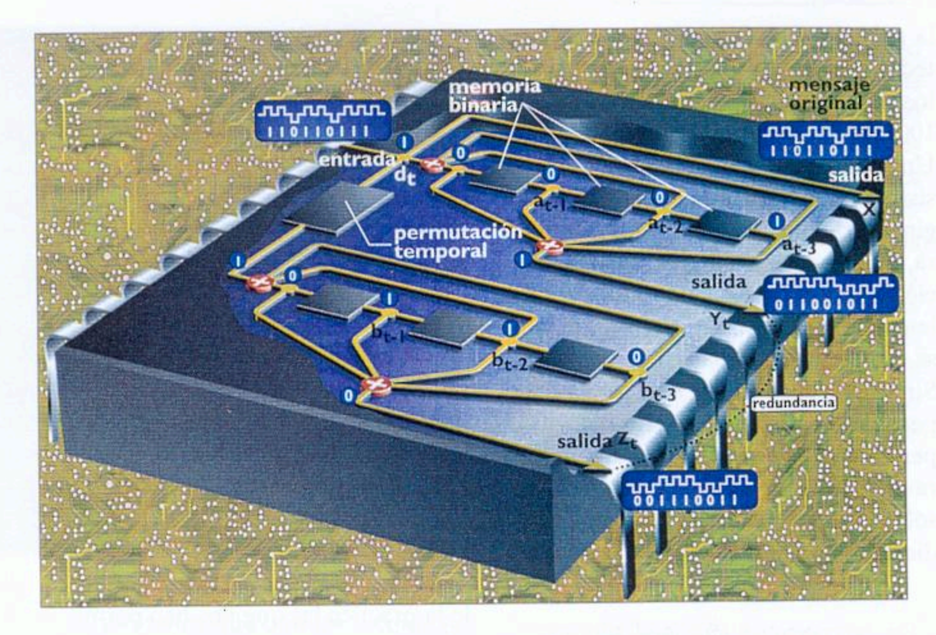


Figura 1. En un circuito integrado, el turbocódigo se presenta como una asociación de dos códigos que transforman el mensaje por medio de sumadores. El subconjunto inferior trabaja con una secuencia permutada. Con estos dos sinónimos y el mensaje original se obtienen tres secuencias para transmitir al descodificador.

Un Ejemplo de Turbocódigo

El mensaje binario {d}, formado por la serie de datos dt, es codificado dos veces: la primera según su orden natural por el primer codificador, arriba (fig. 1) y la segunda según un orden alterado (abajo). Las secuencias redundantes {Y} y {Z} forman dos mensajes sinónimos de la secuencia de entrada {d},que también es transmitida en tanto que secuencia {X}. Por parte del emisor, los datos a transmitir son almacenados en un cierto número de ciclos. Un primer conjunto de circuitos y de memorias proporciona la secuencia binaria no modificada acompañada de otra, pasándola a un sumador que efectúa una suma módulo 2* de la parte de la secuencia almacenada en el registro. El segundo tiene el mismo circuito pero recibe datos en un orden diferente. Las dos secuencias modificadas son formas redundantes de la primera. Por tanto, mediante una adecuada combinación de circuitos pueden regenerar la secuencia de origen. Este código compuesto es absolutamente análogo a la parrilla de un crucigrama. Basta comparar las palabras de la parrilla con el mensaje de origen (serie Xt,anotada {X}) y las definiciones horizontales y verticales con las informaciones redundantes (series de las Yt y Zt, anotadas {Y} y {Z}). Lo mismo puede decirse de las informaciones binarias que llevan las salidas Xt, Yt, Zt del codificador.

Sinónimos. Ya sea para los vehículos espaciales o para los teléfonos móviles, las comunicaciones modernas pasan por una codificación de los mensajes en forma de una serie de símbolos binarios, 0 o 1. En ausencia de palabras, de frases y de una valiosa gramática, la transmisión de datos numéricos tiene que reinventar su lenguaje antierror. El procedimiento más sencillo —se trata ya de una «codificación» del mensaje— consistiría en transmitir dos veces la misma secuencia. Por ejemplo, se enviará 10011010/10011010, conviniendo utilizar paquetes de 16 bits divididos en dos y cuya segunda parte reproduce

la primera. Esa método, aunque detecta los errores, no permite corregirlos. En efecto, ¿cómo elegir entre 10011010 y 10010010?
Una codificación más elaborada consistiría en transmitir la secuencia original acompañada de otra que llegara al mismo resultado. Se trataría en cierto modo de un sinónimo. En el lenguaje corriente, podría emparejarse, por ejemplo, «alba» con «aurora». Sin embargo, aunque esta sucesión permite reparar ciertos errores, no es perfecta: si se recibe «alfa» y «aurora» u «otrora» y «alba», la corrección sólo será posible remitiéndose a un diccionario de sinónimos.

En ausencia de palabras y de una valiosa gramática, la transmisión de datos tiene que inventar su lenguaje anterior

El límite de Shannon. Con un segundo sinónimo, por ejemplo «amanecer», se tendrá una información adicional capaz de favorecer las correcciones. Así pues, pueden añadirse tantos sinónimos como se quiera, pero si los errores de transmisión son demasiado numerosos, esta superposición es incapaz de resolver las ambigüedades. Si, sin ir más allá de dos sinónimos, se recibe «alfa», «otrora» y «amenazar», el asunto se vuelve incomprensible. Por tanto, el poder de corrección tiene un límite. Es el famoso límite de Shannon, que corresponde, por ejemplo, a la situación en que la potencia media de la perturbación es igual a la de la señal útil recibida y se han transmitido el doble de datos que en el mensaje original. Los avances de las técnicas de codificación, constantes hasta 1990, no habían permitido llegar al límite de Shannon. ¿Era este límite inaccesible? Las obras especializadas solían llevar una introducción bastante pesimista sobre la posibilidad de conciliar la teoría y la práctica.
Han sido precisamente las exigencias de la práctica las que nos han permitido elaborar nuevas técnicas de codificación correctoras de errores. Estas técnicas tenían que ser lo bastante sencillas como para permitir la integración en un chip de silicio. La idea consistía en reutilizar informaciones que en general están perdidas y reinyectarlas en el proceso de descodificación, exactamente como un motor con turbocompresor utiliza la presión de los gases de escape para aumentar su eficiencia.

Escribir en líneas y leer en columnas. En 1993, con ocasión de la Conferencia Internacional sobre Comunicaciones de Ginebra, presentamos los sorprendentes resultados de este enfoque.[2] La nueva característica de estos turbocódigos residía en sus prestaciones próximas al límite teórico de Shannon. Perplejos y escépticos, la mayoría de los grandes nombres de la comunidad internacional del cifrado sólo los adoptaron con la prudencia debida. Ideado en paralelo con estos desarrollos teóricos, un circuito integrado ayudó a los más indecisos permitiendo comparar los resultados anunciados con los efectivamente medidos gracias al mismo.[3]
El código utilizado en este circuito era la combinación de dos pequeños códigos convolutivos* unidos por una función de entrelazado. Cada uno de ellos se materializa en un registro de desplazamiento* con tres memorias binarias y sumadores módulo 2 (fig. 1). Una manera simple de realizar un entrelazado consiste en escribir los datos fila a fila y en escribirlos columna a columna (fig. 2).

El código funciona en dos dimensiones como en los crucigramas, donde la redundancia se expresa en las definiciones horizontales y verticales

Así pues, el mensaje binario se codifica dos veces: una en su orden natural y otra en un orden alterado. Al final, este código compuesto es del todo análogo a la parrilla de un crucigrama. Las palabras de la parrilla son comparables al mensaje original, mientras que las definiciones horizontales y verticales son análogas a las informaciones de redundancia introducidas por el código. La metáfora también es válida para la parte receptora: una primera descodificación horizontal permite rellenar ciertas casillas; luego,una segunda, vertical, confirma o cuestiona los resultados anteriores, facilitando el rellenado de las otras casillas. Se vuelve entonces a la descodificación horizontal, con lo que se aportan nuevas letras, y así sucesivamente hasta la convergencia total y la reconstrucción de la parrilla completa.

Dos críticas, un solo autor. Si el turbocodificador tiene una estructura sencilla, el descodificador idóneo es más complejo. Su principio consiste básicamente en el concepto original de infor-

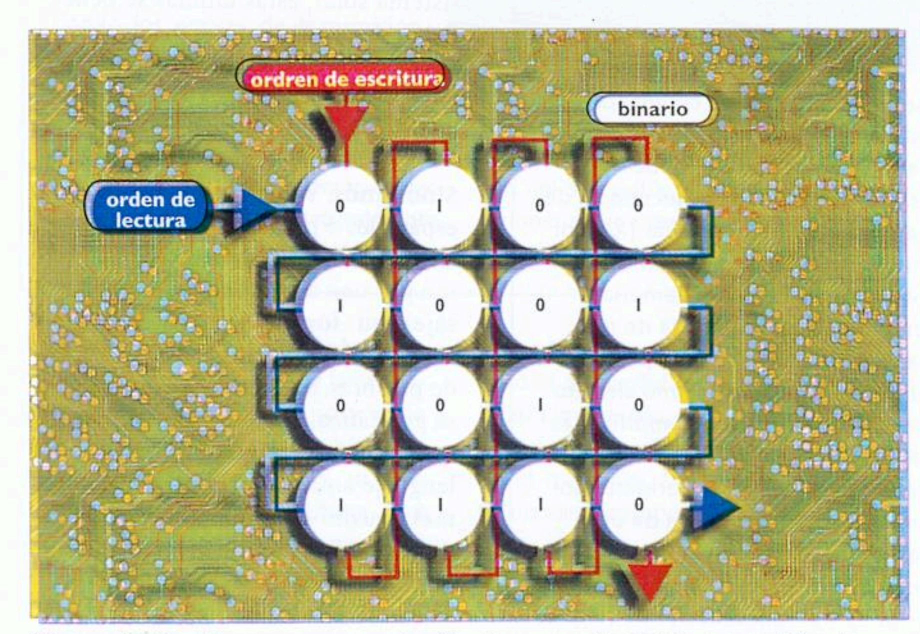
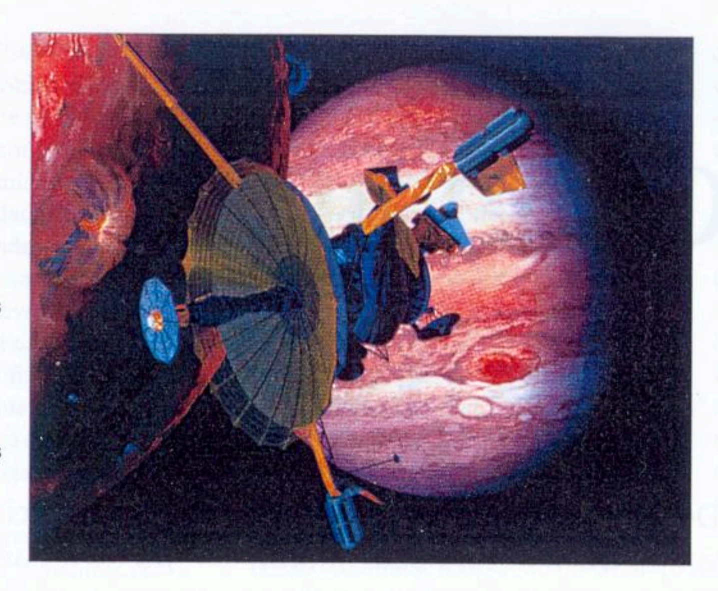


Figura 2. Realizar una permutación temporal consiste en cambiar el orden de lectura respecto al de escritura. Esta operación se realiza en un registro matricial que se rellena columna a columna y se lee fila a fila.



Aceptados por las agencias americana (NASA) y europea (ESA), los turbocódigos se utilizarán en las futuras misiones remotas. También están recomendados en las especificaciones de teléfonos móviles.

mación extrínseca, que podemos ilustrar mediante una situación de la vida cotidiana. Supongamos que queremos comprar un libro reciente, pero que antes queremos formarnos una opinión al respecto.

Leemos dos críticas, precedentes de sendas publicaciones especializadas diferentes con la intención de proceder a una síntesis. En ambos casos la crítica es muy mala, por lo que no compramos el libro.

> La división de un problema complejo en otros dos más pequeños, un concepto habitual en ciencia, permite construir los códigos más eficaces

Pero lo que no sabíamos es que se trataba del mismo crítico, el cual, actuando bajo dos seudónimos diferentes, había vapuleado el libro en ambas publicaciones. Nos habíamos dejado influir por la correlación, pero dos informaciones correlacionadas proceden, total o parcialmente, de la misma fuente. Un órgano de decisión no debe utilizar en modo alguno dos veces la misma información, so pena de error.

Información extrínseca. La información extrínseca es precisamente la que ayuda a escapar de la correlación. Comprende todo lo que no ha sido explotado en la toma de la primera decisión. Así, y volviendo a la imagen del crucigrama, las definiciones horizontales permiten verificar los resultados que dan las columnas y viceversa. La idea original utilizada en la construcción del turbocódigo es, en realidad, un concepto muy extendido en múltiples campos científicos: en vez de tratar de una sola vez un problema complejo —un solo código convolutivo con decenas de memorias— ¿no sería más ventajoso dividir este problema en dos? En tal caso, sabiendo que la complejidad del descodificador crece exponencialmente con el número de memorias del codificador, es mejor tratar sucesivamente dos códigos pequeños que uno solo cuyo número de estados posibles* se cuente por millares o millones.

La utilización juiciosa de este principio permite al «turbodescodificador» ofrecer, gracias a las iteraciones, rendimientos a 0,35 dB* del límite de Shannon. El mejor código «de la competencia» se queda a más de 2,5 dB. Este notable rendimiento ha valido a los turbocódigos ser recomendados en las especificaciones de las futuras generaciones de teléfonos móviles. Aceptados por las agencias estadounidense (NASA) y europea (ESA), se utilizarán en las futuras misiones remotas. Probablemente, la primera será, a partir de 2001, la de *Smart One*, que la ESA se propone enviar a los confines de nuestro sistema solar. ■

Claude Shannon fundó la Teoría de la información y propuso un modelo esquemático lineal de un sistema de comunicación.

CLAUDE BERROU, RAMESH PYNDIAH Y MICHEL JÉZÉQUEL son directores de estudios de la Escuela Nacional Superior de Telecomunicaciones de Bretaña (ENST Bretagne).

ALAIN GLAVIEUX es profesor de la ENST Bretagne

PATRICK ADDE es también profesor de la ENST Bretagne.

(1) C.E. Shannon, «A Mathematical Theory of Communication», *Bell System Technical Journal, 27*, julio y octubre de 1948.

(2) C. Berrou, A. Glavieux y P. Thitimajshima, «Near Shannon limit error-correcting coding and decoding: turbocodes» *in Proc. of ICC, 93*, Geneva, p. 1064, mayo de 1993.

(3) C. Berrou y A. Glavieux, IEEE *Trans. Com., 44*, 1261, 1996.

*LA SUMA MÓDULO 2 es la suma en base dos: 0 + 0 = 0; 1 + 0 = 1; 1 + 1 = 0. Con varias sumas, el resultado es 1 si el número de 1 es impar y 0 si es par.

*LA CODIFICACIÓN CONVOLUTIVA consiste en aplicar a la secuencia de datos una determinada transformación matemática (aquí la suma módulo 2). Por tanto, la secuencia transformada está formada en cierto modo por los datos de origen y un sinónimo. Este tipo de cifrado fue introducido en 1954 por Peter Elias.

*LOS REGISTROS DE DESPLAZAMIENTO son memorias en forma de pila. Un dato que entra empuja los demás a la salida.

*EL NÚMERO DE ESTADOS POSIBLES de un codificador convolutivo es igual al número de combinaciones de «0» y «1» contenidas en las memorias del codificador. Si éste posee «m» memorias (m = 3 en el esquema de la figura 1), el número de combinaciones posibles es 2m.

EL DECIBELIO (DB) es una unidad logarítmica sin dimensiones que expresa la relación entre dos magnitudes —tensión, intensidad o potencia—. Por ejemplo, 2,5 dB corresponde a una razón de 1,8 entre dos potencias.

PARA MÁS INFORMACIÓN:

- Gérar Battail, *Théorie de l'information, application aux techniques de communication*, Masson, 1997.
- Christian Schlegel, *Trellis coding*, IEEE Press, 1997.
- Mariano J. Valderrama Bonnet, *Modelos matemáticos aplicados a las ciencias experimentales*, Ediciones Pirámide, Madrid, 1995.
- C. Heegard y B.S. Wicker, *Turbocoding*, Kluwer Academic Publishers, Boston, MA, 1999.



Los grandes números

Jean-Philippe Bouchaud

De la física a la economía, de la lotería a la sociología, ningún campo de la naturaleza, ningún campo de las actividades humanas escapa a la ley de los grandes números. A gran escala, los comportamientos erráticos se difuminan y hacen una media.

El mundo físico, tal como lo percibimos, obedece a unas pocas leyes simples, como las de la termodinámica, que rigen el comportamiento de gases líquidos y sólidos y los intercambios de energía entre ellos. Elaborada poco a poco a partir del siglo XVII, esta ciencia sólo trata objetos macroscópicos, a escala humana. Pero al principio de nuestro siglo, se impuso la teoría atómica con la idea de que estos objetos están constituidos por un número vertiginoso de partículas microscópicas, generalmente animadas de un movimiento errático e incesante. ¿Por qué milagro un gas, con sus innumerables moléculas puede admitir la descripción sencilla de la termodinámica? ¿Cómo dos fluidos, gases o líquidos, de composición microscópica muy diferente, pueden tener unas propiedades macroscópicas explicables con una misma teoría? La razón de esta repetida simplicidad se debe a una propiedad de una gran generalidad: una acción aleatoria reproducida un gran número de veces tienen consecuencias perfectamente previsibles es la ley de los grandes números, cuyas múltiples facetas intentaremos ilustrar aquí, en particular para comprender el sentido de la frase del gran matemático ruso Andrei Komogortov: «el valor epistemológico de la teoría de las probabilidades se basa en el hecho de que los fenómenos aleatorios engendran a gran escala una regularidad estricta, en la que lo aleatorio, en cierto sentido, ha desaparecido».

Las medidas y las observaciones efectuadas en un gran números de fenómenos aleatorios presentan en promedio una gran regularidad

La desaparición de lo aleatorio se produce a diversos niveles. En primer lugar, y esto es quizá lo más conocido, las medidas y las observaciones efectuadas en un gran número de fenómenos aleatorios presentan en promedio una gran regularidad. A un nivel de observación más fino también se pueden considerar las fluctuaciones aleatorias de las medidas individuales, dicho de otro modo, las desviaciones de la media. Estas desviaciones se describen por medio de leyes probabilísticas, que también son regulares. Todavía más, su comportamiento es ampliamente independiente del detalle de los fenómenos concretos que las originan. Como máximo, se encuentran dos o tres tipos de comportamientos, que se pueden calificar de universales.

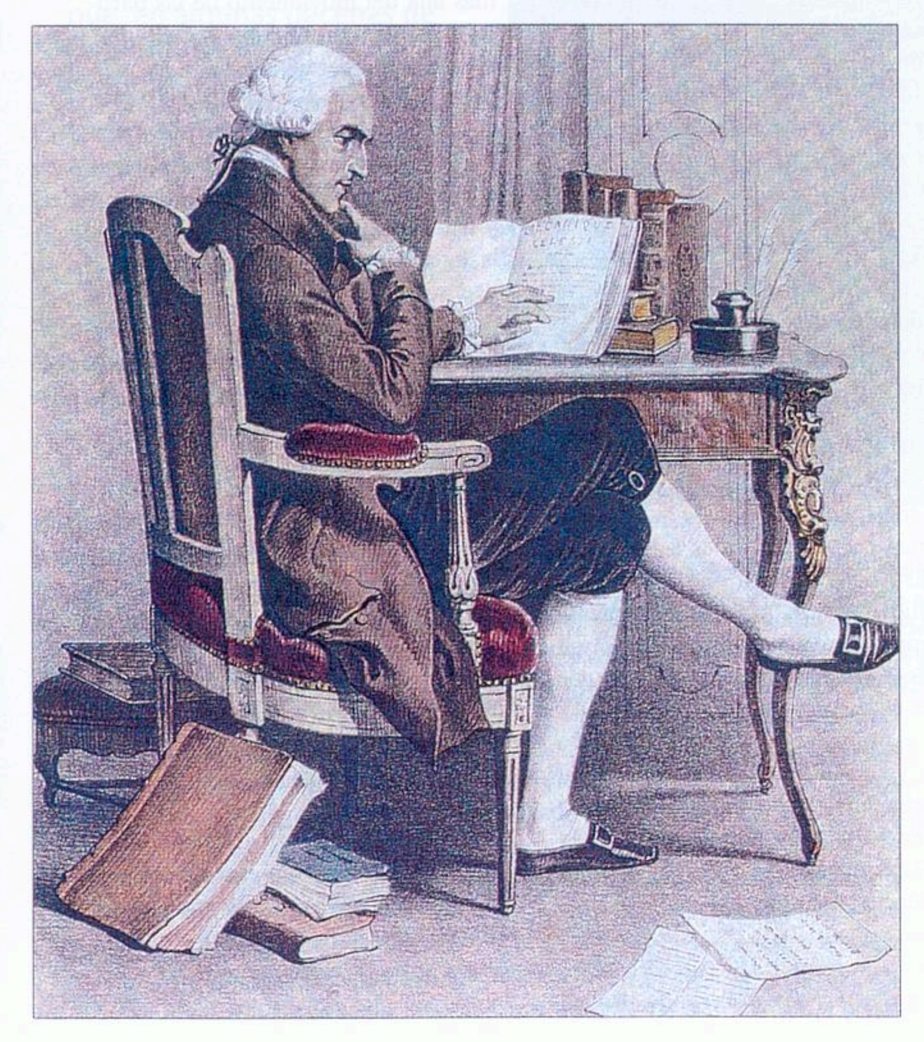
Tanto si se trata de física como de economía, finanzas, sociología u otros campos, esta sorprendente universalidad de comportamientos confiere una gran potencia a los tratamientos matemáticos de las probabilidades y de las estadísticas. Se comprende entonces fácilmente que éstos sean objeto de una intensa actividad investigadora. Los desafíos de estos estudios, tanto si tienen un objeto científico como económico, son considerables: ayudan a optimizar los resultados al conocer mejor la probabilidad de obtenerlos.

Para comprender mejor cómo se difuminan los comportamientos erráticos a gran escala, volvamos al ejemplo de un gas en un recipiente. A la temperatura ambiente, las moléculas que lo componen se desplazan en todos los sentidos a velocidades considerables —varios centenares de metros por segundo— y chocan sin cesar entre ellas y con las paredes. Los choques repetidos de las moléculas contra las paredes son la causa de la presión ejercida por el gas sobre el recipiente. Cada molécula, al rebotar en la pared, le comunica un pequeño impulso que es tan aleatorio como la velocidad que tiene cuando llega. Sin embargo, si las moléculas llegan en gran número a la pared, el impulso total que le confieren por unidad de tiempo casi no fluctúa: los grandes impulsos compensan a los pequeños, ya que se producen todos los casos posibles. Ésta es la forma más sencilla de la ley de los grandes números, claramente observada por Pierre Simon de Laplace hacia el final del siglo XVIII: la suma de N números aleatorios que tienen la misma ley de probabilidad es igual, cuando N es grande, a N veces el valor medio de uno de estos números aleatorios. Si en una bombona de hidrógeno un átomo tiene una energía cinética media e, entonces los N átomos del gas tienen, en total, una energía cinética tanto más próxima a Ne cuanto mayor es N.
Éste es también el principio fundamental de los sondeos de opinión, de los que nos nutrimos diariamente: si el número de individuos encuestados es suficientemente grande, aunque sea moderado respecto a la población de un país en su conjunto, la proporción de «sí» medida es un reflejo fiel de la opinión pública. De la misma manera, cuando se mide la presión de un gas durante un tiempo suficientemente largo, el resultado es único, reproducible, pese a la naturaleza fundamentalmente aleatoria del fenómeno microscópico que la engendra. Además, una ecuación sencilla relaciona la presión con la densidad y la temperatura, es la llamada ley de los gases perfectos establecida en el siglo XVII por Robert Boyle en Inglaterra y Edme Mariotte en Francia, una ley extraordinariamente general y que verifican, al menos aproximadamente, todos los gases, cualquiera que sea su naturaleza.

Hacia 1820 el botánico escocés Robert Brown observó el recorrido errático de un grano de polen en un líquido

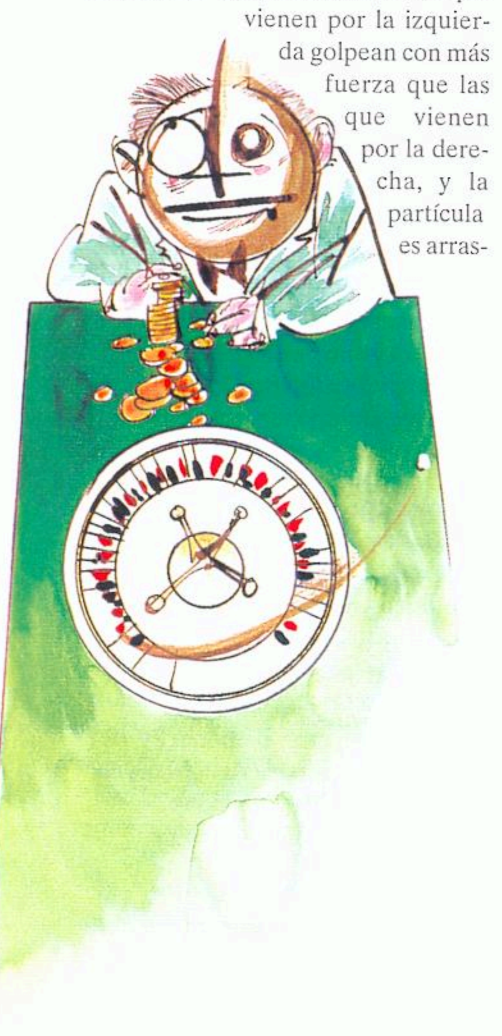
Por lo tanto, no percibimos las fluctuaciones cuando nos interesamos por sistemas macroscópicos. Pero están siempre presentes y basta, para detectarlas, efectuar una medida de la presión durante un tiempo muy corto, o en una porción de pared muy pequeña, de modo que pocas moléculas golpeen la pared. En este caso, la presión se convierte en una magnitud fluctuante, que evoluciona en el tiempo; la densidad y la temperatura ya no bastan para describirla. Se empieza a franquear aquí la frontera difusa entre lo macroscópico y lo microscópico.

Un caso característico en el que las fluctuaciones desempeñan un papel muy importante es el del movimiento browniano. Hacia 1820, el botánico escocés Robert Brown observó, creyendo ver una manifestación de la vida, el recorrido errático de un grano de polen en un líquido. Un grano de polen es una partícula de tamaño microscópico, muy grande comparada con las moléculas del líquido pero demasiado pe-



Se deben a Pierre Simon Laplace (1749-1827) numerosas e importantes trabajos en física astronomía y matemáticas. Su tratado *Teoría analítica de las probabilidades*, publicado en 1812, introduce varias herramientas matemáticas para evaluar las probabilidades de diversos fenómenos naturales. (Foto J.-L. Charmet.)

queña para ser sensible a sus choques individuales. A diferencia de la pared, que sólo sufre los asaltos del líquido por un lado, la partícula browniana se ve afectada por todos lados. Esto significa que el impulso medio comunicado por las moléculas de líquido a la partícula es ahora nulo. Si nos detuviésemos en la ley de los grandes números citada más arriba, se llegaría a la conclusión de que la partícula debería quedar inmóvil, en reposo. Pero como observó Brown, estas partículas se mueven. Cuanto más tiempo pasa, más se alejan de su punto de partida, en un movimiento en zig- zag errático de amplitud creciente. El motor de este movimiento, su razón de ser mencionado en 1900 por Louis Bachelier en Francia e, independientemente, estudiado en detalle por Einstein en 1905, es precisamente la desviación de la media el hecho de que en realidad los choques que recibe la partícula en un instante dado no se compensan totalmente. Por azar, las moléculas que vienen por la izquierda golpean con más fuerza que las que vienen por la derecha, y la partícula es arrastrada por esta fluctuación, mayor por un lado que por el otro; luego, otra fluctuación actuará en otro sentido.





Denis Poisson (1781-1840) fue alumno de Laplace y de Joseph Lagrange. En su obra *Recherches sur la probabilité des jugements...*, publicada en 1837, aparece por primera vez la ley de distribución que lleva su nombre. (Foto J.-L. Chamet).

¿Qué se puede decir de estas fluctuaciones? De modo notable, cuando el número N de acontecimientos considerados es grande, aparece una «ley de los grandes números» ampliada a estas fluctuaciones: es el «teorema del límite central», que nos informa no solamente sobre la media, sino también sobre las desviaciones de la media. Más exactamente, este teorema establece la probabilidad de que se produzca una determinada fluctuación en función de su amplitud, es decir de su desviación respecto a la media. Como era de esperar, las grandes fluctuaciones, que separan mucho a la partícula de su posición media, son menos probables que las pequeñas. Las desviaciones de la media se distribuyen según una ley de probabilidad que es la curva en forma de campana también llamada curva de Gauss (véase el recuadro «Algunas leyes de distribución»). La probabilidad es máxima para una desviación nula y es mucho más pequeña para valores importantes de la desviación. Sin embargo, la anchura de la curva de Gauss, que mide la amplitud de las fluctuaciones probables, crece proporcionalmente a $\sqrt{N}$, la raíz cuadrada del número de acontecimientos, en este caso las colisiones con las moléculas del líquido. Se dice que la posición del grano de polen es una variable aleatoria gaussiana. La partícula tiende a explorar regiones cada vez más alejadas de su posición media, volviendo de vez en cuando cerca de su punto de partida. Señalemos que si estuviese sometida a choques repetidos con efectos no aleatorios, se alejaría proporcionalmente a N, el número de choques, sin regresar nunca.

Este resultado es notable ya que en amplia medida es independiente de los detalles microscópicos, como la naturaleza del líquido circundante y el estado de la superficie de la partícula browniana. Esto es lo que permite, una vez más, describir el fenómeno, a escala del observador, por medio de ecuaciones muy sencillas. Como veremos un poco más adelante, existen sin embargo algunas clases de variables aleatorias que corresponden a fenómenos microscópicos demasiado irregulares para verificar la ley de Gauss.

Desde Laplace y Gauss, se conocen por lo tanto con exactitud las propiedades estadísticas de la desviación respecto de la media en muchas situaciones. Estas propiedades se extienden mucho más allá del movimiento de las partículas brownianas en un líquido. Por ejemplo, el error cometido en los sondeos corresponde, también, a fluctuaciones respecto de la media. Si todos los institutos de sondeo planteasen la misma pregunta en el mismo momento, sus resultados se distribuirían según la ley de Gauss alrededor del valor verdadero, es decir el obtenido encuestando a toda la población. Si se ha encuestado a mil personas sobre un tema, un resultado del 49% de «sí» y del 51% de «no» no es significativo.

¿Por qué? La pequeña diferencia entre los «sí» y los «no» es inferior a las fluctuaciones medias entre los diversos sondeos posibles, que es del orden del 3%. Para alcanzar una precisión del 1% habría que encuestar a una muestra nueve veces mayor; la ley de los grandes números establece en efecto que la incertidumbre relativa, es decir el cociente entre las fluctuaciones N y el valor medio N es $1/\sqrt{N}$.

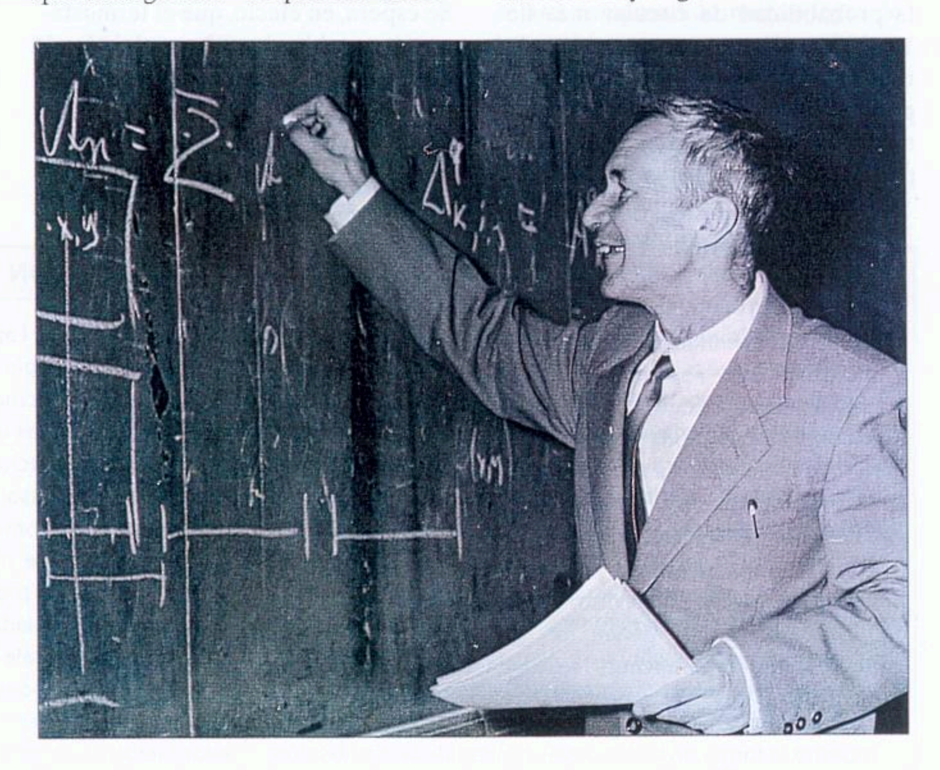
Pero la misma ley de Gauss no puede aplicarse sin algunas precauciones. Sólo es rigurosamente válida para acontecimientos verdaderamente tomados al azar, no correlacionados entre ellos por algún carácter común. Los institu-

tos de sondeo lo saben bien; para optimizar sus previsiones, encuestan a muestras de la población cuidadosamente elegidas, con distribuciones de edad, de categorías socioprofesionales, etc. análogas a las del total del país. Sino los resultados estarían sesgados, y la media obtenida con la muestra estaría más alejada del valor verdadero de lo que establece la ley de Gauss. Por lo tanto, en términos generales, diversificar lleva a acercarse a la ley de Gauss y a reducir las fluctuaciones relativas. Este principio es bien conocido por los buenos gestores: «no hay que poner todos los huevos en la misma cesta». La teoría de la cartera óptima del norteamericano Harry Markovitz, de la Universidad Baruch en Nueva York, laureado con el premio Nobel de economía en 1990, se basa en este principio: promediando la hipótesis de las fluctuaciones gaussianas, esta teoría permite saber cómo diversificar de la mejor manera los activos presentes en una cartera, con objeto de minimizar el riesgo a él asociado.

Es totalmente probable que, en algunas decenas de años, un terremoto produzca diez veces más víctimas que el más mortífero ya sucedido

Lo que está en juego es importante: ¿vienen siempre descritas las fluctuaciones por una ley de Gauss? Si fuese así la incertidumbre sería en cierto sentido controlable ya que un imprevisto realmente catastrófico, como un gran terremoto o un crac bursátil, sería astronómicamente raro. Por ejemplo, un acontecimiento «extremo» cuya desviación de la media fuese de 10 desviaciones estándar (véase el recuadro «Algunas leyes de distribución) sólo tendría una probabilidad 10^{-22} de suceder. Un valor tan pequeño significa que, aunque cada segundo se produjesen diez mil millones de acontecimientos, ¡la edad del Universo apenas sería suficiente para que apareciese un acontecimiento extremo una sola vez! Lamentablemente, está claro que las catástrofes naturales que nos amenazan son mucho más probables: por ejemplo, es totalmente previsible que, en algunas decenas de años, un terremoto produzca diez veces más víctimas que el más mortífero ya sucedido. Este tipo de fenómenos viene descrito por unas variables alteraciones distribuidos según las leyes llamadas de Pareto-Lévy, debido al trabajo premonitorio del economista italiano Vitale Pareto al final del pasado siglo, y del probabilista francés Paul Lévy en los años treinta. Este tipo de variables se caracteriza por su muy gran jerarquía: para una distribución dada (es decir para una determinada realización de estas variables), la mayor de ellas es mucho mayor que la inmediatamente inferior, y así sucesivamente. Una buena ilustración nos la proporciona la distribución de Pareto de los mayores capitales, por ejemplo la cifra de negocios de las mayores empresas europeas: ¡la mayor empresa (Royal Dutch) es dos veces más importante que la segunda! Se pueden hacer constataciones semejantes cuando se examinan las empresas de un determinado sector económico y también las grandes fortunas industriales.

Cuando se consideran las sumas de tales variables, por ejemplo cuando una compañía de seguros hace el balance de los pagos relacionados con las catástrofes naturales, el teorema del límite central que hemos citado, deja de ser válido. De fluctuaciones moderadas en el caso gaussiano, se pasa a fluctuaciones verdaderamente tempestuosas descritas por la ley de Lévy, en las que los acontecimientos más importantes determinan prácticamente del todo la observación. Tal es el caso, por ejemplo, cuando el movimiento errático de una partícula es de naturaleza muy diferente del caso browniano discutido más arriba: la partícula puede progresar a saltos de longitud muy diferente. ¡Aunque los grandes saltos son mucho menos probables que los pequeños, los pocos saltos grandes bastan para explicar a grandes rasgos el camino recorrido! Discutido teóricamente en 1930, este comportamiento sorprendente no fue observado realmente hasta 1990 en conjuntos de micelas gigantes, una especie de moléculas cilíndricas muy largas. [1]



El matemático Andrei Nikolaievitch Kolmogorov (1903-1987) fue una de las grandes figuras de la escuela soviética de probabilidades. Hacia 1930 axiomatizó el cálculo de probabilidades, insertándolo en el marco de la teoría matemática de la medida y de la integración. También fue uno de los creadores de la Teoría de los procesos estocásticos. (Foto (Keystone.)

Lo que realmente notable es que también existe una ley de los grandes números para las fluctuaciones extremas, para los acontecimientos más raros. ¿Cuál es la probabilidad de que el nivel del Nilo alcance un valor particularmente elevado? De nuevo, la respuesta a esta pregunta es, hasta cierto punto, universal, y las leyes posibles son

poco numerosas, como sucede en el caso del teorema del límite central en el que sólo aparecen la ley de Gauss y las leyes de Lévy. La más conocida de ellas es la ley exponencial de Poisson (véase el recuadro «algunas leyes de distribución»). Ésta describe el hecho de que, para que se produzca un acontecimiento raro, se requiere en general una conjunción improbable de varios factores, una sucesión inesperada de golpes de fortuna. Por ejemplo, como se puede constatar analizando las sesiones, la probabilidad de que la bolsa suba N días seguidos decrece aproximadamente como $1/2^N$. En consecuencia cada cuatro años se observan diez días consecutivos de alza, ¡pero veinte días consecutivos se observarán cada cuatro mil años! De la misma manera, en el caso de un automóvil, si la distancia que separa dos pinchazos es, digamos, de 20.000 km en promedio, la probabilidad de circular más de 100.000 km sin tener que cambiar la rueda es de 1/150. Curiosamente, la probabilidad de que, después de un primer pinchazo, se produzca un segundo pinchazo en los 5.000 km siguientes (23%) es más elevada que la probabilidad de un primer pinchazo entre 5.000 y 10.000 km (17%). Se trata de la llamada ley de las series, a menudo invocada por el hombre de la calle.

Las vidrieras de las catedrales fluyen lentamente y engrosan por su parte inferior, cuya transparencia disminuye

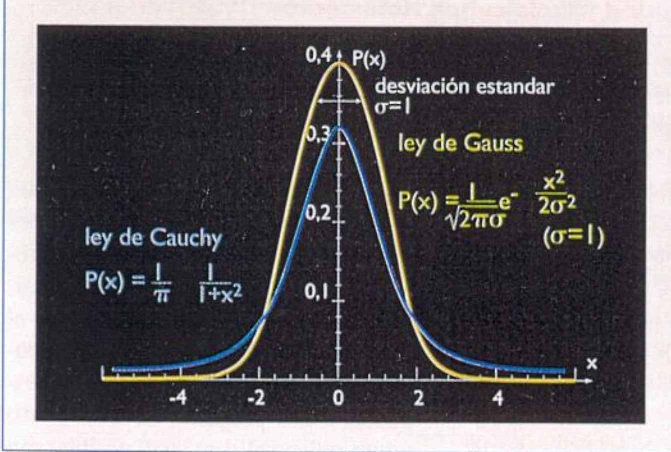
Volviendo a la física, otro ejemplo muy importante en el que aparece esta ley exponencial es la ley de Botzmann que da la probabilidad de que un sistema en contacto con una fuente de calor —un «termostato»— a una temperatura absoluta T adquiera una determinada energía.

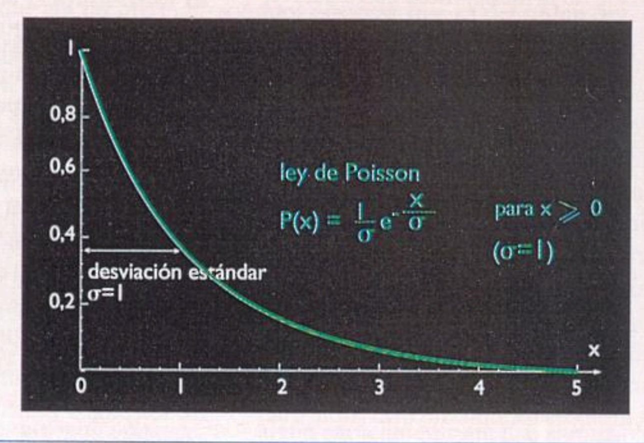
Se espera, en efecto, que el termostato intercambie energía con el sistema. Pero el termostato sólo es capaz de comunicarle energía más allá de ciertos límites. Esto se expresa a través de la probabilidad de que la energía transmitida sea mucho mayor que una energía típica del orden de kT (donde k es la constante de Boltzmann, $k=1{,}38.10^{-23}$ julio/kelvin): es exponencialmente pequeña. Las consecuencias son bien tangibles: materiales utilizados en la vida cotidiana, como los vidrios y algunas aleaciones metálicas destinadas a la aeronáutica, deben sus propiedades al hecho de que no están en su estado de energía más bajo que sería un estado cristalino. La vía que los llevaría a este estado no es directa y exige que se les proporcione energía. Cuando esta energía es algunas decenas de veces superior a la que puede proporcionar, con una probabilidad razonable, el medio ambiente a la temperatura T, el sistema queda atrapado en su configuración temporal durante varios siglos, lo suficiente para que estemos seguros de su perennidad.

Sin embargo, a la escala de siglos, la evolución resulta perceptible. Así las vidrieras de las catedrales fluyen lentamente y engrosan por su parte inferior, cuya transparencia disminuye. Esta dinámica llamada superlenta de

Algunas leyes de distribución

La ley de distribución $P(x)$ de una variable aleatoria x da la frecuencia de aparición de los valores posibles de esta variable. Más exactamente, en el caso de una variable «continua» $P(x)dx$ da la probabilidad de que se observe el valor x de salvo un dx. Por ejemplo, la ley uniforme $P(x)$=constante, describe una situación en la que todos los acontecimientos son equiprobables. El conocimiento de $P(x)$ permite calcular el valor medio x_0 de x. También permite evaluar la anchura de la distribución, es decir el conjunto de valores de x que tienen una frecuencia de aparición apreciable. Esta anchura se llama también desviación estándar (más exactamente la desviación estándar σ se define como σ^2= media de $(x-x_0)^2$. La figura de al lado muestra la forma de las tres leyes clásicas de la distribución: la ley de Gauss (o de Laplace-Gauss, que ya conocía Abraham de Moivre al principio del siglo XVIII), que describe muchos fenómenos con fluctuaciones relativamente moderadas, la ley de Cauchy que es un caso particular de leyes de Pareto-Lévy, con fluctuaciones muy importantes, y la ley exponencial de Poisson. Hay que señalar que existe otra ley de Poisson muy conocida, llamada ley de las pequeñas probabilidades, que rige la probabilidad de ganar en un juego en el que la probabilidad de ganar en cada ensayo una cierta suma g dada en muy pequeña —exactamente si esta probabilidad vale p/N donde N es el número de veces que se juega (se supone que N es grande)—. La probabilidad $P(n)$ de ganar una suma nG vale entonces $P(n)=(P^n e^{-p})/n!$ El valor medio de la ganancia vale pG.





¿Qué es una Variable Aleatoria?

El ejemplo canónico y fundador de una variable aleatoria es el resultado de tirar dados. Por lo demás, *az-zahr* significa dado en árabe. En este caso, la variable aleatoria tiene seis posibles valores con probabilidades iguales a 1/6 (si el dado no está trucado). En este caso, una ley de los grandes números será por ejemplo la correspondiente a la suma de los valores obtenidos arrojando un gran número de datos idénticos. Bien mirado, la noción intuitiva de variable aleatoria plantea dificultades profundas. Por ejemplo, conociendo la posición del dado en la mano del jugador, y las características exactas del movimiento de tirar, se podría predecir el resultado, ya que las ecuaciones de la mecánica son deterministas. ¿De dónde viene el azar, cuál es su fuente última? En el caso del dado, o de un sorteo de la Loto, la respuesta viene de la teoría del «caos»: las ecuaciones del movimiento son desde luego *deterministas*, el menor error, por ínfimo que sea, en la posición del dedo en la mano se amplifica exponencialmente con el tiempo.[1] La incertidumbre sobre el resultado se eleva rápidamente al 100% y *de facto*, ya que no *de jure*, se impone una descripción probabilística o estadística.

Paradójicamente, la realización de un buen algoritmo numérico que permita que un ordenador desempeñe el papel de un casino —una virtud fundamental en un gran número de cálculos (por ejemplo los que recurren al llamado «método Montecarlo»)— no es nada fácil.

los vidrios y de otros sistemas desordenados presenta además algunas particularidades notables que hacen de ellos un objeto de estudio aún muy activo en la actualidad.

A veces, la ley de los grandes números, o una de sus posibilidades, adopta formas más sutiles, y el razonamiento estadístico se introduce de forma más insidiosa. Tomemos el ejemplo de los números primos y su enumeración sistemática: 2,3,5,7,11,13,17,19,23,etc. Mientras estos números son pequeños, su identificación individual tiene un sentido, e incluso puede ser útil. Pero cuando se consideran números muy grandes, el hecho de saber si 13 855 737 o 13 855 741 son primos tiene poco interés por sí mismo; muy pronto, la curiosidad se desplaza de la anécdota individual a la le general: en este caso concreto, la pregunta puede ser cuántos números primos hay en un determinado intervalo, es decir, en el fondo, cuál es la probabilidad de que un número dado sea primo. Esta noción resulta extraña a primera vista, ya que nada es más determinista, menos incierto, que la aritmética. No es la naturaleza del problema la que impone un punto de vista estadístico, sino la magnitud de los números, y la imposibilidad de mantener un registro preciso.

Además, desde hace un siglo se sabe, por ejemplo, cuántos números primos inferiores a un número dado hay en promedio.

El mismo cambio de perspectiva apareció en los años cincuenta, por impulso de los físicos norteamericanos Eugene Wigner y Fred Dyson, en el estudio del espectro de los núcleos atómicos, es decir, de la sucesión de energías permitidas por la mecánica cuántica. Los estados de baja energía se pueden describir en forma relativamente satisfactoria con teorías sencillas. Pero cuando se alcanza el centésimo, estas teorías sencillas son del todo insuficientes, y, a decir verdad, la predicción precisa del centésimo nivel de energía del núcleo no le interesa a nadie.

En cambio, como en el caso de los números primos, es importante saber cuántos estados hay en torno a una energía dada. De forma más importante aún para las aplicaciones físicas, se tienen que conocer las propiedades estadísticas de la diferencia de energía que separa dos estados consecutivos. Sorpresa: las fluctuaciones de esta última cantidad son absolutamente universales, no sólo son independientes del núcleo estudiado, sino que reaparecen idénticas en muchos otros sistemas muy distintos. Un ejemplo de actualidad es el de los materiales conductores desordenados a baja temperatura, cuyas fluctuaciones de conductividad eléctrica vienen caracterizadas por una función que es independiente de la naturaleza detallada del conductor. Relevantemente, desaparece de nuevo el detalle microscópico del problema, y las leyes físicas manifiestan, en detrimento del particularismo, su simplicidad.

Detrás de cada ley de comportamiento macroscópico establecida por los físicos está por lo tanto una ley de los grandes números que justifica la extraordinaria pérdida de información, el abandono de los detalles, que exige una descripción sencilla —accesible a la mente humana—. Recientemente, se han comprendido nuevas leyes de los grandes números, mucho más sutiles, a través del estudio de fenómenos críticos, es decir del comportamiento singular de algunas magnitudes cuando la materia cambia de estado, cuando funde el hielo o cuando un metal se vuelve superconductor. Una vez más, aparecen propiedades universales pese a la diversidad microscópica de los sistemas estudiados. Aplicada a las sociedades humanas y a los individuos, esta regularidad absoluta de la que hablaba Kolmogorov produce vértigo: ¿son nuestras acciones individuales algo más que la confirmación de una tendencia general que nos supera? ■

JEAN-PHILIPPE BOUCHAUD es ingeniero de la CEA (Comisaría de la energía Atómica), en Francia y trabaja en el Servicio de Física del Estado Condensado en el Centro de Estudios de Saclay, cerca de París. Fue responsable de investigaciones en el CNRS entre 1985 y 1992.

Mundo Científico ha publicado:
J.P.Bouchaud «Los "vuelos de Lévy"o la difusión no browniana», nº113, mayo,1991.

Para Más Información:

- W.Feller, *An introduction to probability theory*, Wiley, 1977.
- B.V. Gnedenko y A.N. Kolmogorov, *Limit distributions for sums of independent variables*, Addison-Wesley, 1954
- P. Lévy, *Théorie de l'addition des nombres aléatoires*, Gauthiers-Villars, 1954
- E.W. Montroll y M.F. Shlesinger, «The wonderful world of random walks» en *Non equilibrium phenomena II, Studies in statistical mechanics*, J.L. Lebowitz y E.W. Montroll (eds.), North Holland, 1984.
- E. Vincent *et al., en Recent progress in random magnets*, D. H. Ryan (ed.), World Scientific, 1992.
- *Liquids, freezing and glass transition*, Les Houches 1989, J.-P. Hansen *et al* (eds.), North Holland,1991.
- O. Bohigas y M.J. Giannoni en *Mathematical and computational methods in nuclear physics*, J.S. Dehasa *et al.*, (eds.), Springer-Verlag, 1983
- John Allen Paulus, *El hombre anumérico*, Tusquets Editores, Colección Metatemas, Barcelona, 1999.



Los códigos correctores

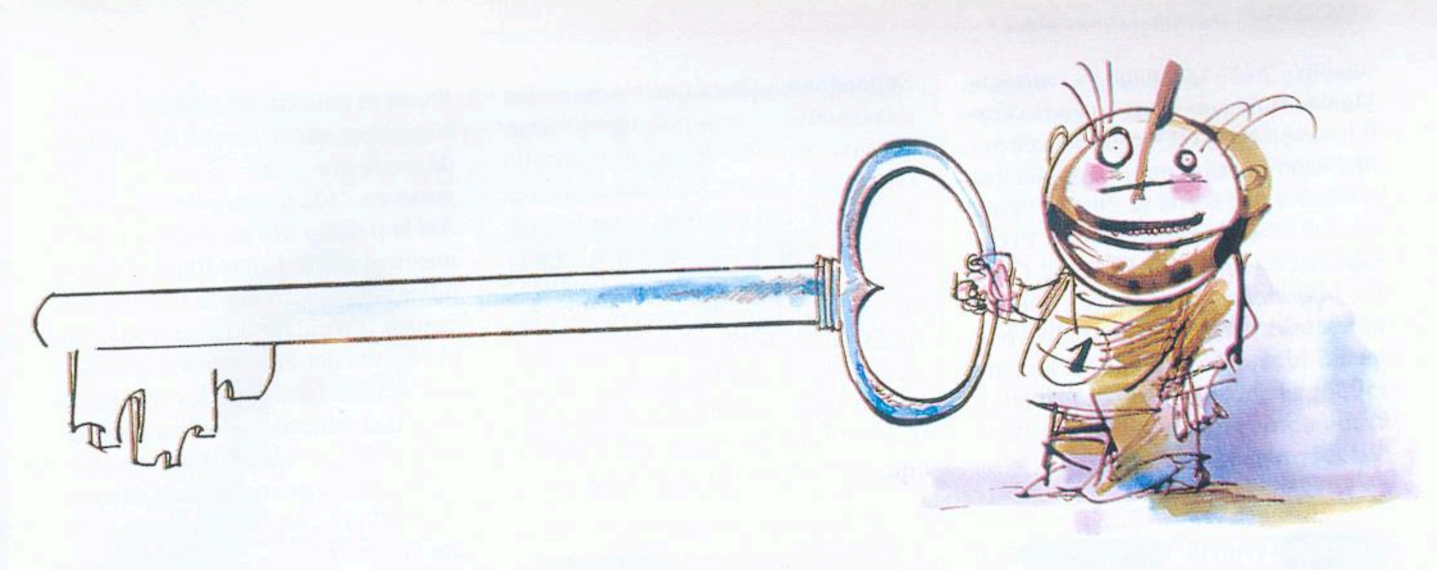
Gilles Lachaud y Serge Vladut

Actualmente, representar la información en forma de una sucesión de números se ha convertido en algo totalmente corriente. Pero es necesario que los códigos empleados sean capaces de transmitir los mensajes con eficacia y de corregir los inevitables errores. Un desafío que intentan superar los especialistas en teoría de los números.

Comunicar es transmitir mensajes. Para ello, el hombre utiliza los sistemas de signos y los lenguajes más diversos. La lengua natural recurre a frases construidas con palabras, a su vez constituidas por letras. Casi todo el mundo conoce también los jeroglíficos, los ideogramas chinos, los pictogramas de la guía Michelin, las notas musicales, etc. Estos sistemas de signos que permiten representar o transmitir la información, son los llamados códigos. Codificar un mensaje es pasar del significado al significante. Ya en sus *Pensamientos*, Blaise Pascal escribía un texto premonitorio en muchos aspectos :*«Cifra tiene dos sentidos, uno claro y el otro del que se dice que el sentido está oculto. Las lenguas son cifrados en los que no son las letras las que se transforman en letras, sino las palabras en palabras, de modo que una lengua desconocida es indescifrable».*

La historia y la tradición nos han legado la mayoría de los códigos que utilizamos. Pero otros han sido especialmente concebidos con fines precisos. Así, el objeto de la criptografía ea concebir códigos, a ser posible inviolables, que aseguren la confidencialidad de los mensajes correspondientes. No abordaremos este problema, por lo demás muy interesante. También se puede desear solamente concebir códigos eficaces que permitan garantizar una transmisión de los mensajes lo más fiel posible. La importancia de este objetivo es evidente a la vista de la

Figura 1. Uno de los códigos binarios más antiguos es sin duda el constituido por los 64 hexagramas del Yi-King chino, construidos yuxtaponiendo seis símbolos binarios (yin: — y yang: — —).



revolución —numérica— que afecta en nuestros días a la informática y las telecomunicaciones (redes mundiales como Internet, teléfono, transmisiones por satélite y por fibras ópticas, etc.). Se comprende por lo tanto que se esté desarrollando una carrera desenfrenada para poder transmitir mucha información, muy rápidamente, con una tasa fiable de errores y con el menor coste.

El estudio de la codificación se inició verdaderamente hace unos cincuenta años, cuando el norteamericano Claude E. Shannon fundó la teoría matemática de la información. Sus trabajos sucedían a los de Ralph W.L. Hartley y Harry Nyquist sobre la transmisión de señales, al final de los años veinte. Shannon demostró que existen códigos optimales tanto en lo que concierne a la corrección de eventuales errores de transmisión como en lo que concierne a la comprensión de la información. Lamentablemente, sus teoremas no proporcionaban un método que permitiese construir explícitamente los códigos más eficaces. Faltaba por lo tanto encontrar unos códigos eficaces, pero el camino no estaba trazado del todo. Aunque ha habido un gran número de adelantos, la búsqueda continúa activamente en nuestros días, en particular la relativa a los «códigos correctores de errores».
En esta investigación de códigos eficaces, las matemáticas son un instrumento esencial. Incluso más: la teoría de los números desempeña un papel clave. ¿Por qué? Porque, aunque se puede codificar la información con la ayuda de no importa qué símbolos, una codificación eficaz tiene que hacer intervenir relaciones entre los símbolos transmitidos. La forma más cómoda de describir tales relaciones es asimilar los símbolos utilizados a números enteros y utilizar todos los recursos de la aritmética, el álgebra y el análisis para encontrar y representar las operaciones adecuadas a las que se tienen que someter los números- símbolos.

El yi-King es un discurso sobre el sentido atribuido a las «palabras» de tres a seis letras binarias

El ejemplo que tenemos probablemente todos en mente es la codificación binaria utilizada en los ordenadores y más en general por la tecnología numérica: este código utiliza dos símbolos 0 y 1, que corresponden a dos estados físicos posibles de los elementos del dispositivo técnico por el que circula la información. Se habría podido llamar A y B a estos dos símbolos, o + y –. Por lo demás, la historia de la codificación binaria empieza pronto, ya que uno de los cinco clásicos de la China antigua, el Yi-King, utiliza los símbolos *yin* (–) y *yang* (– –) (fig. 1). El yi-King es un discurso sobre el sentido atribuido a las «palabras» de tres o seis letras binarias, y las matemáticas chinas utilizaban corrientemente estos conceptos. En cierto sentido, el *ying* y el *yang* son equivalente a 0 y 1, como ya había señalado Leibnitz. Queda el hecho de que, generalmente, la representación en forma de números enteros 0 o 1 es la más cómoda, como se ha dicho, en la medida en que hay que considerar la realización de operaciones con estos símbolos.

Veamos ahora cómo se define un código en términos más matemáticos. ¿Qué se necesita para representar la información? Primer ingrediente: un alfabeto finito **F**, el decir un conjunto finito de símbolos. Se convierte entonces el mensaje a transmitir en una sucesión de «palabras», siendo cada palabra una secuencia de símbolos, de longitud fija k. Dicho de otro modo, una palabra u es una secuencia $u_1u_2...u_k$, donde cada u_i es un elemento del alfabeto **F**. Las palabras son por lo tanto elementos del «conjunto producto» $\mathbf{F}\times\mathbf{F}\times...\times\mathbf{F}=\mathbf{F}^k$ (el lenguaje natural emplea palabras cuya longitud no es fija, pero dejaremos de lado esta posibilidad). Un conjunto de palabras define un «vocabulario». Los matemáticos llaman generalmente codificación, o código, a una aplicación de un vocabulario A en un vocabulario B, es decir una aplicación que hace corresponder a toda palabra de A una única palabra de B (además para que sea posible la descodificación, dos palabras diferentes de A tienen que estar asociadas a dos palabras diferentes de B). Por abuso de lenguaje, los vocabularios A y B —cuyas alfabetos y longitud de las palabras no son necesariamente idénticos— se califican también a menudo de códigos.
He aquí algunos ejemplos para fijar las ideas. Aunque los alfabetos utilizados para codificar los mensajes son de naturaleza muy diversa —ideogramas, notas musicales, letras, etc.— el más sencillo es sin duda el conjunto de dos elementos $\mathbf{F}_2=\{0,1\}$. Un elemento de $\mathbf{F}_2$ es un *símbolo binario* también llamado *bit* (del inglés *binary digit*). Este alfabeto es el que utiliza la tecnología informática. Según la longitud elegida se pueden construir diversos tipos de palabras. Las palabras que tienen 4 bits se llaman números *hexadecima-*

les; hay 2^4=16 (a menudo se representan de forma más concisa con la cifras 0,1....,9,A,B,C,D,E,F del sistema de numeración hexadecimal). Las palabras formadas por 8 bits se llaman bytes. Existen 2^8 =256 bytes diferentes. El código ASCII (*American Standard Code for Information Interchange*), que representa los caracteres tipográficos por medio de bytes comprende por los tanto 256 palabras. Éste es el código utilizado por los ficheros de tipo «TEXT» de los ordenadores.

Supongamos ahora que los mensajes a transmitir se han puesto en forma de una sucesión de palabras constituidas por *k* símbolos tomados de un alfabeto **F**. Las palabras pasarán a un canal de comunicación (línea telefónica, onda de radio, cinta magnética, etc.) (fig. 2). Pero diversas fuentes de ruido (ruido de fondo de los aparatos electrónicos, cualidad mediocre de las señales emitidas, fotografías transmitidas borrosas, etc.) hacen que el canal no sea nunca perfecto;

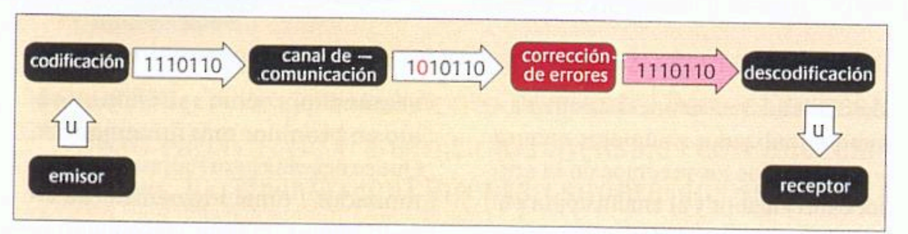


Figura 2. En la transmisión de un mensaje interviene una codificación, por ejemplo en forma de una secuencia de 0 y 1. Al pasar al canal de comunicación, fuentes de ruido muy diversas pueden hacer que esta secuencia sufra modificaciones inoportunas. Si el código está bien construido, estos errores se pueden detectar y corregir. Éste es el objeto de los «códigos correctores de errores».

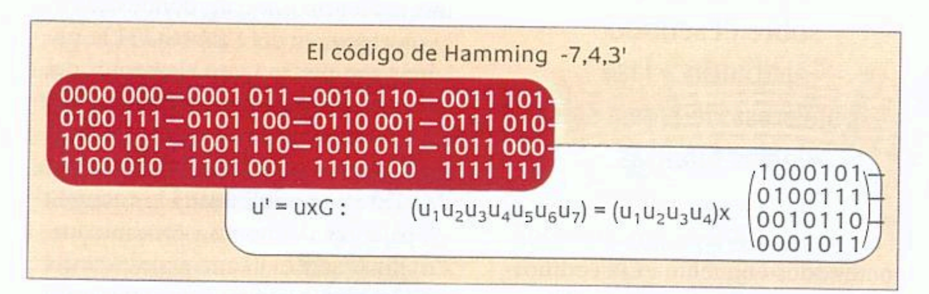


Figura 3. Este código de Hamming transforma palabras de 4 bits en palabras de 7 bits, añadiéndoles tres símbolos de control calculados a partir de los cuatro símbolos de información iniciales. Esta transformación de una palabra u=$(u_1u_2u_3u_4)$ en otra palabra u=$(u_1u_2u_3u_4u_5u_6u_7)$ se puede describir por medio de una matriz G. En este código, el número mínimo de símbolos que distinguen dos palabras diferentes es 3, lo que permite detectar y corregir un eventual error en uno de los 7 bits de la palabra.

El código genético constituye otro ejemplo. En las células vivas, los aminoácidos se sintetizan a partir del ácido ribonucleico (RNA). Cada aminoácido está codificado en el RNA por una unidad de tres nucloótidos. El RNA está constituido por cuatro tipos de nucleótidos que son la adenina (A), la citosina (C), la guanina (G)y el uracilo (U). La codificación de los aminiácidos en el RNA consiste por lo tanto en «palabras» de tres «letras», o codones, construidas a partir del alfabeto A,C,G,U. El codón UAC, por ejemplo corresponde al triptófano. Existen 4^3=64 *codones* diferentes, pero algunos de ellos codifican para el mismo aminoácido, mientras que otros no contienen información genética.

siempre hay un riesgo, más o menos pequeño, de que algunos símbolos del mensaje se emitan o se transmitan deforma errónea. Si se considera el alfabeto binario {0,1}, la recepción de un «1» en vez de un «0» puede se catastrófica para la interpretación del mensaje. ¿Cómo hacer que la transmisión de la información sea menos vulnerable a este tipo de errores casi inevitables? Un aforismo muy conocido afirma que la repetición es la base de la enseñanza.
Los códigos correctores de errores utilizan una estrategia análoga. En vez de enviar simplemente palabras de *k* símbolos al canal, se alargan añadiendo a cada una de ellas un cierto número de *símbolos de control*. Un ejemplo sencillo, en el caso de un alfabeto binario, consiste en añadir un «bit de paridad», de modo que el número de «1» que figuran en cada nueva palabra sea par. Así la palabra 010 se alargará a 0101, mientras que 011 dará 0110. Si se modifica un bit durante la transmisión, cambia la paridad y este procedimiento permite detectar un error (pero no permite corregirlo). En los códigos de identidad bancaria se utiliza un sistema análogo, se añade una letra clave en el número de cuenta para detectar un eventual error de transmisión. Dicho de otro modo, se envían mensajes *redundantes:* no sólo se transmite la información principal, sino que el mensaje contiene una información sobre el propio mensaje.

En términos generales, un código corrector de errores hace corresponder a cada palabra de *k* símbolos una palabra más larga, por ejemplo de *n* símbolos. Se trata por lo tanto de una codificación que a cada palabra $u_1u_2...u_k$ del conjunto $\mathbf{F}^k$, le hace corresponder una palabra $u_1u_2...u_k...u_n$ del conjunto $\mathbf{F}^n$, donde *n* es un entero mayor que *k*. Hay que señalar aquí que, aunque se utilicen todas las palabras de $\mathbf{F}^k$, no sucede lo mismo con todas las palabras de $\mathbf{F}^n$.

¿Cómo elaborar un código corrector de errores? Ya hemos subrayado el interés que presenta un alfabeto constituido por números. En efecto, es indispensable poder efectuar operaciones con los símbolos, por ejemplo para *calcular* los símbolos redundantes o de control a partir de los otros. Las operaciones más sencillas son la suma, la resta, la multiplicación y la división. Un conjunto en el que están definidas estas cuatro operaciones se llama *cuerpo*. (Para una definición más precisa, véase en este ejmplar el artículo sobre los números *p*-ádicos.) En el caso de la codificación, se necesita un alfabeto **F** constituido por un número finito de símbolos: por consiguiente **F** tiene que ser un cuerpo finito. Se comprende por lo tanto que los especialistas en códigos intentes explorar al máximo esta rama de las matemáticas puras que es la aritmética de los cuerpos finitos. Además, si **F** es un cuerpo finito, cada palabra de $\mathbf{F}^k$ (respectivamente de $\mathbf{F}^n$) se puede considerar como un vector en un espacio de dimensión *k* (respecti-

vamente n), y los símbolos u_1,u_2, etc. son entonces las componentes del vector en cuestión. Esto permite utilizar todos los recursos del álgebra lineal. En muchos códigos, los símbolos de control que se añaden para alargar las palabras dependen linealmente de los símbolos de información $u_1u_2...u_k$; en este caso se habla de códigos lineales. Se pueden considerar una generalización del procedimiento mencionado más arriba, consistente en añadir un bit de paridad.

En los códigos lineales, la transformación que a cada palabra de $\mathbf{F}^k$ le hace corresponder una palabra de $\mathbf{F}^n$, se puede representar de forma muy cómoda por medio de una matriz G, que es una especie de tabla de números. Supongamos por ejemplo que se quieren transmitir palabras de 4 bits $u_1u_2u_3u_4$; el código de Hamming —por el nombre del ingeniero norteamericano R.W. Hamming que inventó estos códigos en los años cincuenta— añade tres símbolos redundantes dados por las relaciones $u_5 = u_1+u_2+u_3$, $u_6 = u_2+u_3+u_4$, $u_7 = u_1+u_2+u_4$ (con las reglas aritméticas 0+0=0,0+1=1,1+1=0). Así, la palabra 0011 se transforma en 0011101, 1100 se convierte en 1100010, etc. En este código lineal, la transformación de una palabra u en una palabra más larga u' se puede escribir $u'=uG$ donde G es una matriz muy sencilla en este caso, que contiene cuatro filas y siete columnas (fig. 3).

Se puede demostrar que este código lineal puede corregir un error producido en uno de los bits de la palabra enviada. En general, para corregir errores en la recepción, lo mejor es que dos palabras distintas tengan muchas coordenadas diferentes. Si, por ejemplo, dos palabras cualesquiera difieren en al menos tres coordenadas, se puede corregir un error de transmisión en una sola letra: si la palabra recibida no pertenece al «vocabulario», la palabra enviada es la única palabra del código que difiere de la palabra recibida en una sola letra. En consecuencia, para recuperar la palabra inicial, se utiliza el principio de descodificación de *máxima verosimilitud*, que supone que ha cometido un mínimo de errores. En otras palabras se estima que la ocurrencia de varios errores simultáneos en una misma palabra es poco probable, tanto menos probable cuanto más elevado es el número de errores posible.

Cuerpos Finitos y Curvas sobre un Cuerpo Finito



Evariste Galois (1811-1832)
Este retrato se debe a su hermano Alfred Galois, que dedicó un verdadero culto a la memoria de su hermano. (Foto J.-L. Charmet)

En la elaboración de códigos eficaces, la teoría de los cuerpos finitos interviene de forma fundamental (en efecto, el alfabeto de los símbolos es un cuerpo finito). Un cuerpo es esquemáticamente, un conjunto de elementos en el que la suma, la resta, la multiplicación y la división «funcionan» correctamente. Entre los cuerpos más conocidos figura el de los números reales, es decir el de los números usuales (racionales e irracionales): contiene un número infinito de elementos. Un cuerpo finito, por su parte, sólo contiene un número finito de elementos. Un ejemplo de los más sencillos es el cuerpo de los elementos F_2={0,1}, con la multiplicación habitual y la regla de suma 0+0=0,0+1=1+0=1, y 1+1=0 (se trata de la «suma de módulo 2»): se hace la suma clásica, pero sólo se considera el resto de la división por dos).

La teoría de los cuerpos finitos le debe a Evaristo Galois, matemático francés. Por ello los cuerpos finitos también se llaman cuerpos de Galois y los anglosajones denotan GF(q)(por «Galois field») al cuerpo finito F_q que contiene q elementos. La teoría de Galois lleva en especial a una clasificación completa de los cuerpos finitos. He aquí algunos de sus principales resultados:

- Todo cuerpo finito F_q contiene un subcuerpo F_p siendo p un número primo. El número de elementos de F_q es necesariamente una potencia de p:$q=p^n$ donde n es un entero positivo.
- Reciprocamente, para todo $q=p^n$, donde n es un entero positivo y p un número primo, existe un único cuerpo F_q que tiene q elementos.
- Todo cuerpo F_p donde p es primo se puede identificar con el conjunto {0,1,2...p-1} en el que la suma y la multiplicación se hacen «módulo p», es decir, considerando únicamente el resto de la división por p (por ejemplo, para p=5,3+4=2 y 3x3=4).

Se puede explorar en qué se convierten diferentes objetos matemáticos cuando forman parte de un cuerpo finito. Así, de todas las ecuaciones consideradas en un cuerpo finito, la categoría más estudiada es la de las ecuaciones con dos variables y, más en general, los sistemas en los que hay una variable más que ecuaciones. En geometría cartesiana con números reales estas ecuaciones son bien conocidas: curvas «planas» en el caso de las ecuaciones con dos variables como $y=x^2$ (una parábola), superficies alabeadas en el espacio. Cuando los parámetros y las variables de la ecuación son elementos de un cuerpo finito, se dice que la ecuación define una curva en el cuerpo finito considerado. Se trata de curvas algebraicas, ya que sus ecuaciones vienen siempre dadas por polinomios. En efecto, en un cuerpo finito, todas las funciones son polinomios, lo que simplifica enormemente los cálculos: no hay ni senos ni cosenos (esto se debe a que para todo elemento x de un cuerpo finito, se tiene $x^q=x$ donde q es el número de elementos del cuerpo). Uno de los resultados más importantes concierne al número de puntos de una curva algebraica en un cuerpo finito, es decir al número de soluciones del correspondiente sistema de ecuaciones. El matemático francés Anfré Weil, uno de los fundadores del grupo Bourbaki, demostró en 1940 que el número N de puntos de la curva verifica la desigualdad $N \leq (q+1+2g\sqrt{q}$ donde q es el número de elementos del cuerpo considerado y g es el «género») de la curva, un número que mide su complejidad). La generalización de esta desigualdad le valió a Pierre Deligne la medalla Fields en 1978.



Por tanto, el número mínimo de símbolos diferentes de una palabra a otra del código es un parámetro importante, que no siempre es fácil de calcular. Designémoslos d. Cuanto más grande es estre número, más capaz es el código de corregir los errores. Para que el código sea capaz de detectar y rectificar un bit erróneo, es necesario que d sea al menos igual a tres. Otros dos parámetros que caracterizan un código corrector de errores son la longitud n de cada palabra (Símbolos de información iniciales + símbolos de control) y el Número k de símbolos de información. Se habla de códigos $[n,k,d]$. El ejemplo de código de Hamming expuesto anteriormente es un código [7,4,3]. Los parámetros n,k, y d permiten definir una tasa de transmisión por el cociente k/n que mide la velocidad de transmisión y también su coste. En efecto, cuanto más se han alargado las palabras más pequeño es y más lenta resulta la transmisión. Por su parte, el cociente d/n mide la fiabilidad de la transmisión. Por ejemplo, aunque para detectar un error puede bastar un símbolo de control, se necesitan por lo menos dos símbolos más para corregirlo. Éste es el caso de la codificación de triple repetición, que transforma 0 en 000 y 1 en 111; en este caso se descodifica decidiendo que el símbolo emitido es el que más se repite (por ejemplo, si se recibe la palabra 101, se deduce que ha enviado 111). Pero este código [3,1,2] es muy costoso: la tasa de trasmisión es un símbolo recibido por cada tres emitidos, es decir 1/3.

Los trabajos de Shannon en los años cincuenta, combinados con los del matemático soviético Rom Varshamov y del norteamericano E.N. Gilbert hacia 1952-1953, demostraron que se podían construir códigos con una tasa de transmisión óptima respecto a la fiabilidad. Lamentablemente, estos resultados no indican cómo obtenerlos.

Los códigos de Hamming son más bien mediocres en relación con los criterios definidos por Varshamov y Gilbert. Pero posteriormente los matemáticos han elaborado muchos otros, de mayor rendimiento, por medio de métodos muy diversos que generalmente recurren a conceptos elaborados de la teoría de los números, y en particular de la aritmética en los cuerpos finitos. Así sucede, por ejemplo, con los códigos de Reed-Solomon (inventados en 1960 por I.S. Reed y G. Solomon en el MIT), que tienen muchas aplicaciones. Citemos entre otras los discos audionuméricos (los CD) y la telemetría de los satélites civiles, en los que intervienen códigos de Reed-Solomon basados en los cuerpos finitos $\mathbf{F}_{64}$ y $\mathbf{F}_{256}$ respectivamente (64 y 256 designan el número de elementos que tiene el cuerpo).

Cómo construir un código de Goppa en el caso de una curva plana en un cuerpo finito

Hace quince años, el matemático ruso Valery Goppa descubrió que por medio de curvas sobre cuerpos finitos se pueden obtener códigos eficaces, que constituyen una amplia generalización de los códigos de Reed-Solomon clásicos. ¿Qué es una curva en un cuerpo finito? Una curva ordinaria en el plano está definida por una ecuación con dos variables x e y, donde x e y representan las coordenadas de un punto. La ecuación $x^2+y^2=1$, por ejemplo, representa una circunferencia de radio 1. Habitualmente, las variables x e y, así como los parámetros que caracterizan la ecuación pueden tener *a priori* cualquier valor real. Pero, en el caso de una curva en un cuerpo finito, los parámetros y variables son únicamente los elementos del cuerpo finito considerado; hablar de curva es ligera-

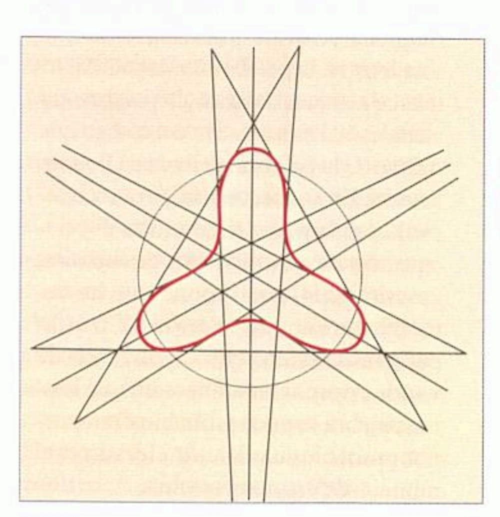
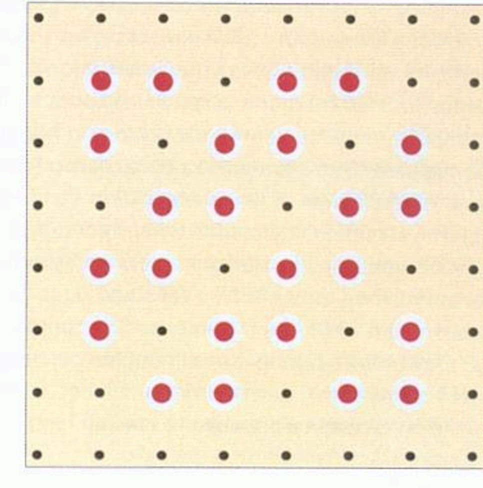


Figura 3. A la izquierda, la curva de Klein dibujada en el plano euclídeo (dibujo según F. Klein)

A la derecha la curva de Klein en el «plano» definido por el cuerpo finito de 8 elementos. Este «plano» sólo contiene 8x8=64 puntos. El número de puntos de la curva de Klein es de 24 y permiten construir un código corrector de errores eficaz que transforma palabras de longitud 3 en palabras de longitud 24 (dibujo según I. Duursma).



Codificación y Ecuaciones Diofánticas

Codificar con números significa no sólo que los números sirven de alfabeto para la codificación, sino también que se utilizan los métodos de la teoría de los números para desarrollar los códigos. Entre las herramientas principales figuran las ecuaciones diofánticas, es decir, las ecuaciones polinómicas con varias variables y de coeficientes enteros. Estas ecuaciones deben su nombre a Diofanto de Alejandría (325-410). Una de las más sencillas es la «ecuación de Fermat» $x^n+y^n=1$, donde n es un entero positivo fijado y x e y números racionales a determinar. Para $n=2$, esta ecuación representa una circunferencia y su solución en números racionales se conoce desde la antigüedad. Pierre de Fermat (1601-1665) enunció sin demostración que para n mayor que 2, la ecuación no posee soluciones racionales no nulas. Este célebre «gran teorema de Fermat» se ha demostrado hace tan sólo unos pocos meses, en una hazaña sin precedentes.[I,II] Al igual que la ecuación de Klein, la de Fermat se utiliza para la construcción de códigos, que puede tener muchas soluciones cuando se buscan en un cuerpo finito. Mencionemos a este respecto que el cálculo del número de soluciones de la ecuación de Fermat con $n=3$ y en un cuerpo finito F_p es uno de los resultados más profundos del matemático Carl- Friedrich Gauss (1777-1855).

mente abusivo, ya que número de puntos (x,y)distintos que verifican la ecuación de la curva es en este caso finito (cuerpo finito obliga...).
He aquí, a modo de orientación, cómo construir un código de Goppa en el caso de una curva plana en un cuerpo finito. Una curva plana en un cuerpo finito. Una curva de este tipo se puede representar siempre por medio de una ecuación $f(x,y)=0$, donde f es un polinomio de dos variables, cuyos coeficientes pertenecen al cuerpo finito elegido. Los puntos de la curva son los pares $P_1=(x_1,y_1)$, $P_2=(x_2,y_2)$,.... $P_n=(x_n,y_n)$, que cumplen $f(x,y)=0$ (n es el número de puntos de la curva). Formar un código lineal equivale a definir la matriz G que transforma una palabra u en $u'=uG$. Esto se hace de la siguiente manera. Se fija un entero positivo m y se elige una «base» $\{G_1, G_2,...G_k\}$ de los polinomios con dos variables de grado inferior o igual a m (se demuestra que $k=m\ (m+1)/2$. El término «base» significa que todo polinomio de grado inferior o igual a m se escribe de forma única como una combinación lineal de los polinomios G_1, G_2,...G_k (por ejemplo, los polinomios clásicos de una variable y grado 2 tienen la forma ax^2+bx+c; una base muy sencilla y muy cómoda es la constituida por los tres polinomios x^2, x y 1).

La etapa siguiente es la construcción de una matriz *G* de n columnas y k filas, cuyo elemento G_{ij} situado en la fila i y en la columna j viene dado por $G_{ij}=G_i(p_j)$(dicho de otro modo, G_{ij} es el valor del polinomio G calculado en el punto $P_j=(x_j,y_j)$ de la curva. La matriz G define un código de Goppa en la curva f, considerada en un cuerpo finito. Se trata de un código $[n,k,d]$, en el que d se puede calcular fácilmente (es una de las ventajas del método): en efecto, se puede demostrar que d, el número mínimo de símbolos que distinguen dos palabras diferentes, es superior o igual a n-deg(f), donde def(f) es el grado del polinomio $f(x,y)$, que es la ecuación de la curva. Se ve también que cuanto más elevado es el número n de puntos de la curva, más largo es el código que se obtiene.

Actualmente se dispone de técnicas de descodificación extraordinariamente rápidas, sencillas y eficaces

Esta construcción se puede aplicar a la curva de Klein $x^3\ y+y^3+x=0$ considerada por ejemplo en el cuerpo finito $\mathbf{F}_8$ de 8 elementos y obtener un código de parámetros [24,3,21] (fig. 4). En este caso, palabras de longitud 3 se codifican en palabras de longitud 24. La tasa de transmisión no es muy buena, ya que las palabras son muy largas; en cambio, la «distancia» mínima d entre dos palabras distintas es 21, un valor muy elevado que permite corregir hasta 10 errores simultáneos en una palabra de 21 símbolos. La invención de estos códigos es muy reciente. Sus distancias mínimas son grandes y por lo tanto pueden resultar útiles en las transmisiones muy perturbadas, por ejemplo las efectuadas por satélite. Esta construcción de códigos de Goppa, también llamados códigos geométrico-algebraicos, se revela totalmente explícita y efectiva. Por ejemplo, Thomas Zink, de la Universidad de Berlín, Michel Tsfasman y uno de nosotros, Serge Vladut, entonces en la Universidad de Moscú, demostraron en 1981 que existe toda una familia de códigos de Goppa en el cuerpo finito **F**q de q elementos que se pueden construir explícitamente y que funcionan mejor que el óptimo definido por Varshamov y Gilbert cuando q es superior o igual a 49.

Estos resultados son un paso importante hacia la elaboración de los códigos de máxima eficacia predichos por la teoría de Shannon. Permiten, utilizando construcciones más clásicas, obtener códigos binarios totalmente eficaces. Además, desde hace poco se sabe descodificar de manera explícita los códigos construidos de esta forma. El muy renombrado IEEE (*Institute of Electrical and Electronics Engineers*) norteamericano ha concedido tres premios a los trabajos efectuados sobre esta materia. Hasta ahora, los códigos obtenidos a partir de ecuaciones algebraicas todavía no se han utilizado mucho, ya que los algoritmos de descodificación no eran suficientemente efectivos. Pero se han realizado grandes esfuerzos por parte de diversos grupos de investigadores de diferentes países. De modo que actualmente se dispone de técnicas de descodificación extraordinariamente rápidas, sencillas y eficaces. Esperemos que dentro de dos o tres decenios, los códigos geométrico-algrebraicos sustituyan a los códigos clásicos actuales.

GILLES LACHAUD es director de investigación en el CNRS. Dirige el equipo de investigación «Aritmética y Teoría de la información» en el Laboratorio de Matemáticas Discretas de Marseile-Luminy, en Francia.

SERGE VLADUT es director de investigación en el Instituto de los Problemas de la Transmisión de la Información de Moscú. También es director de investigación asociado al CNRS.

Mundo Científico ha publicado:
C. Goldstein, «El teorema de Femat» nº146, mayo, 1994.
C. Goldstein, «La conjetura de Femat ya es un teorema» nº160, setiembre, 1995.

Cálculo simbólico y automatización

Dominique Duval

Una vez automatizado el cálculo numérico, la informática acometió la automatización de operaciones matemáticas más abstractas. Los programas de cálculo formal, ya de fácil acceso, ahorran a los científicos horas e incluso años de trabajo. Pero hay que saber utilizarlos.

Para calcular, por ejemplo, el volumen de una bola de 5 cm de radio, hay que efectuar un cálculo «numérico». Basta conocer la fórmula $V = 4\pi r^3/3$, sustituir en ella el símbolo π por 3,14159... y el símbolo r por 5 para obtener un valor (aproximado) del volumen: $V = 523{,}59...\ cm^3$. Supongamos ahora que queremos expresar el volumen de una bola en función de su área S, *a priori* desconocida. Sabiendo que S vale $4\pi r^2$, es suficiente expresar el radio r en función de S y sustituir esta expresión en la fórmula $V = 4\pi r^3/3$. Se encuentra fácilmente $V = S^{3/2}(6\sqrt{\pi})$. El cálculo realizado es un ejemplo de cálculo «simbólico», llamado también «formal» (y a veces analítico o «algebraico»).

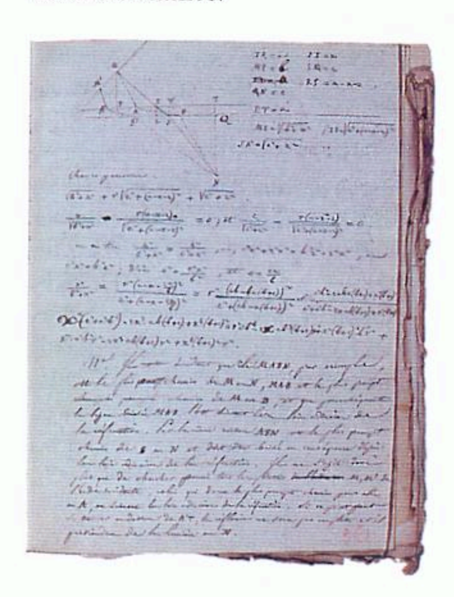
Un cálculo exacto. Estos ejemplos muy simples ilustran los dos principales tipos de cálculo que deben realizar los científicos. En la práctica, dichos cálculos pueden efectuarse con lápiz o papel o con ordenador. Actualmente, el recurso al ordenador —o a una calculadora— para el cálculo numérico es cosa corriente. Como se trata, en general, de cálculos aproximados (ya que el número de decimales está fijado) se ha llegado a decir que los ordenadores «calculan mal».[1] En realidad, nada prohíbe a los ordenadores «calcular bien» y manejar símbolos y expresiones matemáticas. Pero la existencia de programas encargados de estas tareas es mucho más reciente y menos conocida que la automatización de los cálculos numéricos.

Los programas —se habla también de «sistemas»— de cálculo formal hicieron su aparición hace más de treinta años: en 1958, en Estados Unidos, John McCarthy utilizaba el lenguaje Lisp para obtener derivadas* de expresiones simbólicas. Los físicos, que han de habérselas a menudo con cálculos largos, tuvieron un papel importante en el desarrollo de los primeros sistemas. En Francia, muchos informáticos y matemáticos descubrieron el cálculo formal en los años 1980 a raíz de la creación en el CNRS de una estructura de investigación sobre el tema dirigida por Daniel Lazard.

Pero sólo en los últimos años el uso de los sistemas de cálculo formal ha empezado a difundirse más allá del círculo de los especialistas. Han contribuido distintos factores: culturales (algunos jóvenes han tenido ocasión de utilizar este software), técnicos (progreso de los ordenadores en lo tocante a rapidez y capacidad de memoria), comerciales (una publicidad más cuidada) y científicos (aumento considerable de la gama de problemas resolubles).

Reservados durante largo tiempo a un pequeño mundo de especialistas, los programas de cálculo formal empiezan a penetrar en las empresas

Pero estos programas, debido tal vez a la dificultad de su empleo, son muy poco utilizados todavía en las empresas. Esta situación podría modificarse rápidamente introduciendo una iniciación al cálculo formal en la mayoría de las enseñanzas científicas (véase recuadro «Enseñanza y cálculo formal») e integrando dicho cálculo en las calculadoras y los programas de cálculo científico.



Unas fórmulas simplificadas. El alcance del cálculo formal es considerable. En las distintas ciencias, las matemáticas distan de limitarse a determinar valores numéricos aproximados. La simplificación de las fórmulas, esto es, su escritura en forma fácilmente legible, es crucial para captar su significado. El científico suele tener que resolver, o al menos que estudiar y manejar, ecuaciones dadas en forma no numérica. Incluso en el caso de los cálculos numéricos hay que

saber plantear el problema del modo más conveniente, lo cual exige un tratamiento formal previo. Estas manipulaciones matemáticas de todo tipo requieren en general mucho tiempo e ingenio, con un riesgo no despreciable de error. Los programas de cálculo simbólico pueden ser, en este aspecto, unos auxiliares preciosos.

Ciento veintiocho páginas de cálculo. Se habla a menudo del cálculo de la posición de la Luna en función del tiempo que realizó el astrónomo francés Charles Delaunay. Delaunay publicó su resultado en 1860 y 1867 (en dos volúmenes) pero tardó unos veinte años en elaborarlo y verificarlo. La fórmula ocupa 128 páginas de su libro *La Teoría del movimiento de la Luna*. En 1970, el cálculo fue realizado por ordenador en menos de un día y, como propina, se detectó un pequeño error en el resultado de Delaunay.

Los cálculos simbólicos largos, complicados y fastidiosos no son exclusivos de la mecánica celeste. También abundan en matemáticas, química e ingeniería, así como en casi todos los sectores de la física.

Consideremos, por ejemplo, el momento magnético del electrón. Las medidas experimentales dan un valor muy preciso. Pero para determinar con idéntica precisión el valor propuesto por la teoría, y poder comparar ambos, hay que realizar un cálculo simbólico demasiado complicado, difícil de imaginar sin un programa de cálculo formal.

En un registro ligeramente distinto, los sistemas de cálculo simbólico prestan servicios en criptografía y codificación (para transacciones bancarias, comunicaciones e inclusos discos compactos). Estos campos necesitan realizar cálculos en el marco de unas estructuras matemáticas bastante abstractas llamadas «cuerpos finitos». Ahora bien, contrariamente al cálculo numérico, que no sabe qué hacer con ellas, el cálculo formal las trata con bastante facilidad.

Lo mismo ocurre con el manejo automático de los programas informáticos. Un programa, sea cual sea el lenguaje en que está escrito, es un dato simbólico que un sistema de cálculo formal puede leer, estudiar y transformar, lo cual permite ganar tiempo y evitar errores.

Manejo de símbolos. La especificidad del cálculo formal reside en su capacidad para manejar símbolos. He aquí dos ejemplos absolutamente elementales a los que volveremos:

1)Resolver la ecuación $ax = b$, donde

Muchos cálculos simbólicos largos y fastidiosos pueden encargarse ya a un ordenador. Los manuscritos llenos de manipulaciones matemáticas, como los que aquí se reproducen, ¿están en vías de extinción? (A la izquierda, cálculos de óptica de un manuscrito de Augustin Fresnel; a la derecha, páginas de Einstein en las que figuran cálculos de relatividad general; arriba, cálculos de mecánica celeste por Urbain Le Verrier). (Foto biblioteca del Instituto/ J.-L. Charmet, archivos magnum Foto y en Haut, Observatorio de París)

x es la incógnita. Numéricamente, sólo es posible resolver la ecuación para valores dados de a y b. Los sistemas de cálculo formal, por su parte, manejan directamente los símbolos a y b y dan la solución $x = b/a$.

2) Evaluar la expresión $(1 + x - 1)/x$. Para un programa de cálculo simbólico, esta expresión siempre es igual a 1. Un programa de cálculo numérico, en cambio, sólo puede evaluarla para un valor dado de x y responde 0 cuando el valor de x es lo bastante pequeño.

El cálculo numérico realiza operaciones con números «flotantes», esto es, con aproximaciones de números reales. Así, 1/3 se escribe 0,3333333333 (el número de decimales depende de la precisión del programa); por consiguiente, 3 multiplicado por 1/3 no vale exactamente 1. En cambio, el cálculo formal maneja los números de manera exacta: 1/3 permanece en esta forma y el programa sabe las reglas de cálculos de fracciones, por lo que, para él, 3 multiplicado por 1/3 vale exactamente 1. Asimismo $\sqrt{2}$ se mantiene en esta forma en vez de representarse en forma decimal. El programa conoce las reglas del cálculo con radicales, por lo que el cuadrado de $\sqrt{2}$ es igual a 2, sin aproximación alguna. Análogamente, el número π se considera como un símbolo dotado de ciertas propiedades, como sen $\pi = 0$.



Antaño, el cálculo formal estaba reservado a ordenadores relativamente pesados y potentes. Han aparecido ahora en el mercado unas calculadoras compactas que ofrecen unas posibilidades de cálculo simbólico relativamente extensas. La calculadora de la foto es capaz de desarrollar expresiones algebraicas, factorizarlas, calcular el desarrollo en serie de potencias, etc. (Foto ©Texas Instruments)

Enseñanza y Cálculo Formal

En Francia, el cálculo formal se había utilizado hasta ahora en la enseñanza a un nivel puramente experimental. En la enseñanza secundaria, al existir una versión en francés, se solía optar por el programa Drive, que funciona con ordenadores poco potentes. En el primer ciclo de enseñanza superior, la introducción al cálculo formal era ocasional e inconexa. Se asiste ahora, con la aparición de las primeras calculadoras simbólicas, a una importante entrada en la enseñanza secundaria. En la enseñanza postsecundaria hay una introducción al uso de un programa de cálculo formal en ciertas clases preparatorias a las grandes escuelas.

Estos programas subrayan los aspectos positivos: «Se trata de acostumbrar a los estudiantes a utilizar programas que ayudan al razonamiento mediante una confrontación rápida y cómoda de las hipótesis con los resultados y permiten [...] aligerar la parte de cálculo sistemático en beneficio de la intuición matemática y del sentido físico.»

En matemáticas, el aspecto experimental puede extenderse a nociones más complicadas que antes. En física y química, el cálculo formal constituye una herramienta para analizar resultados: por ejemplo, la rápida resolución de ecuaciones diferenciales y la representación de soluciones para distintos valores de los parámetros permiten comprender mejor ciertos fenómenos.

Pero hay otra cara de la moneda. Aunque se suprimen los cálculos fastidiosos en favor de una mejor comprensión de las nociones, se corre también el riesgo de eliminar, en los exámenes, la validación de las competencias simples.

Obsevemos, por último, que una vez más la generalización del empleo del cálculo formal va a ser demasiado brutal de cara a la formación de los enseñantes. Ahora bien, los aspectos pedagógicos son fundamentales para un buen uso de una herramienta que, como las calculadoras numéricas o gráficas, requiere un espíritu crítico. Los programas de cálculo formal ponen en entredicho el contenido de los programas de matemáticas y de las competencias exigibles. Esto ocurría ya con las calculadoras numéricas y gráficas a un nivel más bajo (operaciones, construcciones de curvas). Pero las calculadoras formales tienen que ver con todos los conocimientos algorítmicos básicos en matemáticas. La reflexión sobre la necesaria reforma de la enseñanza de las matemáticas la llevan a cabo enseñantes aislados o estructuras como el IREM (Institut de recherche sur l'enseignement des mathématiques). La brusca democratización del cálculo formal pone a los enseñantes contra las cuerdas. ¿Sabrán estar a la altura de las circunstancias?

Anne Bellido
Maître de conférences de la Universidad de Limoges y doctora adjunta del IREM de Limoges.

Otra diferencia del cálculo formal respecto al numérico es que en aquél el tamaño de los números enteros no está limitado. Por ejemplo, 50! vale 30414093201713378043612608166064768844377641568960512000000000000. Un número como ese no es aceptable en un sistema de cálculo numérico porque rebasa con mucho la capacidad de memoria disponible. En tal caso, hay que contentarse con un valor aproximado del tipo $50! \approx 0{,}30414093 \times 10^{65}$. De hecho, semejante valor puede darlo también un sistema de cálculo formal; además, el usuario puede imponer la precisión deseada, es decir, el número de cifras significativas.

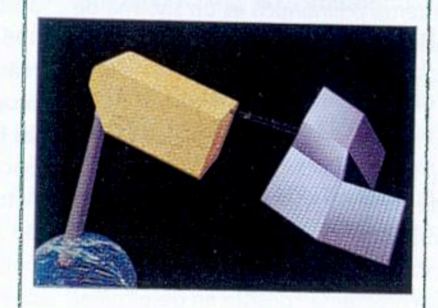
Gracias a su exactitud, el cálculo formal permite estudiar problemas numéricamente inestables como la sucesión definida por: $a_0 = 11/2, a_1 = 61/11$, y $a_{n+1} = 111 - 1130/a_n + 3000/(a_n a_{n-1})$ para $n \geq 1$ (un ejemplo debido a Jean-Michel Muller, de la Escuela Normal Superior de Lyon[11]). Numéricamente, cuando se calculan los términos uno tras otro, la

sucesión parece converger rápidamente hacia 100 cualquiera que sea la precisión; con una precisión de 8 cifras, $a_n = 100$ cuando $n \geq 14$, y con una precisión de 16 cifras, $a_n = 100$ cuando $n \geq 27$. Pero este resultado es falso, pues es posible demostrar rigurosamente que el límite de la sucesión es igual a 6. Con un programa de cálculo formal se puede calcular para todo n el valor de a_n (se trata de un número racional, cociente de dos enteros) y luego deducir un valor aproximado. Se ve entonces que la sucesión converge hacia 6. Se verifica, por ejemplo, que $5{,}9999999 \leq a_{100} \leq 6$. En otras palabras, el cálculo numérico da pronto un resultado falso ($a_{100} \approx 100$), mientras que el cálculo formal, tras un tiempo indudablemente más largo, obtiene el resultado correcto ($a_{100} \approx 6$). Las habilidades de los sistemas de cálculo formal no acaban aquí ni mucho menos. Estos programas permiten desarrollar expresiones algebraicas como ésta: $(a + b)(a - b)$, o, al contrario, factorizarlas (fig. 1). También saben factorizar los polinomios* de una o más variables, las fracciones racionales*, las funciones usuales* (exponencial, logarítmica, seno, etc.) y las «funciones especiales» (funciones de Bessel, hipergeométricas, etc.) y las matrices (multiplicación, determinante, matriz inversa, etc.)



El astrónomo y matemático francés Charles-Eugène Delaunay tardó unos veinte años en realizar y verificar un cálculo preciso de la posición de la Luna en función del tiempo. En 1970, un ordenador logró realizar este cálculo formal en menos de un día... (Portrait Peint por M. Sevestre, foto observatorio de París)

Sistemas Mecánicos: El Programa James



El programa JAMES, elaborado por Aerospatiale y la sociedad Simulog, modeliza, analiza y simula sistemas mecánicos con varias articulaciones. El diseño y la construcción de vehículos espaciales, como satélites de observación o de comunicaciones, pero también de vehículos automóviles terrestres, requieren este tipo de simulación. JAMES ha sido utilizado, entre otras cosas, para modelizar y simular la apertura de los generadores solares de los satélites Spot 4 y Helios (véase ilustración).

Las etapas típicas de la simulación de un programa con JAMES son: la definición del sistema mecánico y de su configuración, la determinación de las ecuaciones que describen el movimiento, el análisis y eventualmente la modificación de dichas ecuaciones, y la construcción del programa destinado a resolver numéricamente el problema. JAMES utiliza el sistema de cálculo formal Maple para efectuar los cálculos simbólicos y manipular las fórmulas. (Documento Simulog)

Saben resolver muchos tipos de ecuaciones y de sistemas de ecuaciones, calcular las derivadas de las funciones usuales y de sus combinaciones, desarrollar una función en serie de potencias de la variable (por ejemplo $\cos x = 1 - x^2/2 + x^4/24 - x^6/720 + ...$), hallar primitivas, etc. Las manipulaciones matemáticas así automatizadas abarcan aproximadamente todas las que aprenden los estudiantes en los dos primeros años de universidad y muchas más.

Robótica y sistemas polinómicos: el programa GB

Un sistema polinómico es un sistema de ecuaciones de la forma:

$P_1(x_1, x_2,..., x_n) = 0$

...

...

$P_m(x_1, x_2, ..., x_n) = 0$

donde los $Pi(x_1, ..., x_n)$ son polinomios. Los sistemas polinómicos intervienen en muchos campos y especialmente en robótica. Un problema consiste en determinar la posición de la plataforma conociendo la configuración robot (en particular, el número de patas y su longitud). Ello equivale a resolver un problema polinómico: las ilustraciones de esta página representan tres de las 40 posiciones posibles para un mismo robot de seis patas. Para resolver un problema polinómico, el austriaco Bruno Buchberger propuso en 1965 asociarle otro sistema polinómico más simple (base de Gröbner). También propuso un algoritmo para el cálculo de las bases de Gröbner, que es simple pero poco eficaz.

Actualmente, siguen calculándose bases de Gröbner para resolver los sistemas polinómicos. Por sus trabajos sobre estas cuestiones y sobre sus aplicaciones a la robótica, Jean-Charles Faugère, investigador de la Universidad de Paris-VI, recibió en 1995 el premio Seymour-Cray. Faugère ideó un programa, llamado GB, que utiliza algoritmos más sofisticados que los de Buchberger. La implantación de GB se realiza en un lenguaje de bajo nivel (el lenguaje C++) y es cuidadosamente optimizada, por lo que la ganancia de eficacia es notable. Por ejemplo, el sistema polinómico de cuatro variables:

$a + b + c + d = 0$

$ab + bc + cd + da = 0$

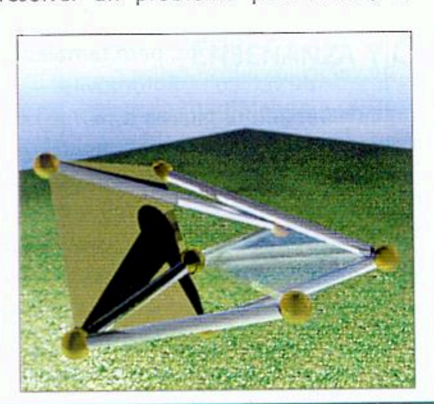
$abc + bcd + cda + dab = 0$

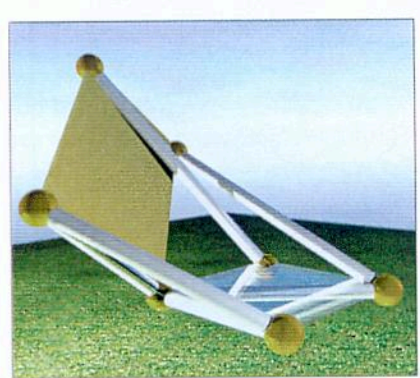
$abcd = 0$

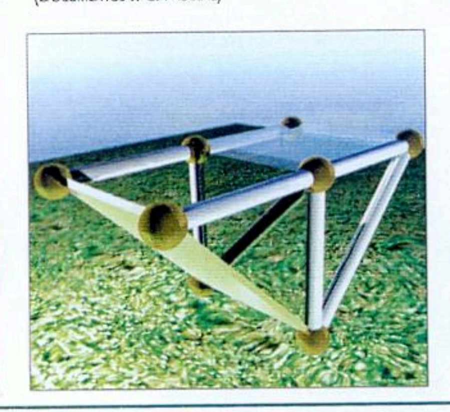
es fácil de resolver a mano, pero hace unos diez años ningún programa de cálculo formal sabía tratar el sistema análogo de cinco variables.

El programa GB lo resuelve actualmente en menos de un segundo y también resuelve con éxito, en una jornada de ordenador, el sistema análogo de ocho variables (pero el resultado ocupa cientos de megaoctetos).

(Documentos J.-C. Faugère)







Un programa por disciplina. La lista completa de las posibilidades ofrecidas sería muy larga. Es posible hacerse una idea de ellas consultando el manual de ayuda «en línea» de un sistema de cálculo formal. Esta lista depende poco del programa considerado, al menos para los sistemas generalistas como Macsyma, Reduce, Maple, Mathematica, Axiom, MuPad, etc. Hay también muchos sistemas que suministran algoritmos especializados en tal o cual campo, por ejemplo, Schooschip para la física de partículas elementales, Camal para la mecánica celeste y la relatividad, Magma y Gap para la teoría de grupos (véanse también los recuadros).

La mayoría de los problemas concretos requieren un tratamiento en dos o tres etapas: en primer lugar un manejo formal de expresiones o incluso de programas, seguida de una resolución formal (a menudo parcial) que puede depender de parámetros; viene luego un cálculo numérico que completa la resolución para valores fijados de muchos parámetros; resulta necesaria, por último, una visualización gráfica de los resultados. Como estas tres fases recurren a programas que nada tiene que ver entre sí, los problemas de coherencia constituyen a veces un auténtico rompecabezas. Por ello, los sistemas de cálculo formal ofrecen además posibilidades numéricas (traducción a lenguajes como Fortran o C) y gráficas (trazado y animación de curvas o superficies). Se trata de auténticos sistemas de *cálculo científico* en sentido amplio, que integran todos sus aspectos. El cuadro que acaba de esbozarse podría dar a entender que los programas de cálculo formal son omnipotentes e infalibles y que basta pulsar unas pocas teclas para obtener la respuesta buscada. Nada más lejos de la realidad. Es muy probable que los sistemas ideales de cálculo formal nunca vean la luz. Los programas actuales son interactivos, baratos y de fácil empleo para problemas que, por su naturaleza y tamaño, se acercan a problemas tipo. No obstante, para sacar pleno partido de sus numerosos recursos hay que tener una idea de su modo de funcionamiento, de sus límites y también de los límites intrínsecos de los problemas considerados.

Un gran éxito del cálculo simbólico: el manejo de polinomios que figuran en todos los cálculos

Elementos básicos. Ante todo, hay que saber que las posibilidades que ofrece un programa de cálculo formal suelen ocultar resultados matemáticos sofisticados, un estudio muy fino de los algoritmos y de su implantación y una elección juiciosa de la representación de los datos. Además, como es fácil de comprender, uno de los *leitmotiv* es la obtención de mejores tiempos de cálculo. Consideraremos aquí dos ejemplos, el primero referente al cálculo polinomial y el se-

gundo a la integración de funciones. Los polinomios constituyen elementos básicos para el cálculo formal; se construyen a partir de números, símbolos y operaciones de adición, sustracción y multiplicación. Los polinomios figuran implícita o explícitamente en todos los cálculos, por lo que es imprescindible saber manejarlos lo más eficazmente posible. Una operación fundamental, que interviene por ejemplo en la simplificación de una función racional, es el cálculo del máximo común divisor (*mcd*) de dos polinomios. Recordemos que el *mcd* de dos polinomios $F(t)$ y $G(t)$ es el polinomio de mayor grado $Q(t)$ que factoriza a la vez $F(t)$ y $G(t)$ (es decir, tal que $F(t) = F_1(t)Q(t)$ y $G(t) = G_1(t)Q(t)$, donde F_1 y G_1 son polinomios). Para determinar este *mcd* existe un algoritmo simple cuyo principio ya era conocido por Euclides. Se trata de una generalización elemental del algoritmo simple que aprendimos en la escuela primaria para efectuar divisiones. A pesar de ello, la determinación del *mcd* es objeto de intensas investigaciones en el cálculo formal. ¿Por qué? Consideremos los dos polinomios:

$F(t) = t^8 + t^6 - 3t^4 - 3t^3 + 8t^2 + 2t - 5$

y $G(t) = 3t^6 + 5t^4 - 4t^2 - 9t + 21$.

La aplicación directa del algoritmo de Euclides lleva al resultado: $Q(t) = mcd(F, G) = 1288744821/543589225$.

En Estados Unidos se ha elaborado un algoritmo sofisticado de cálculo formal para determinar el MCD de dos polinomios

Máximo común divisor. Se comprueba que $Q(t)$ no depende de t. Ello significa que los polinomios *F* y *G* son primos entre sí (no tienen ningún factor común de grado mayor o igual a 1). El resultado tiene 10 + 9 = 19 cifras pese a equivaler a $mcd(F, G) = 1$, ya que el *mcd* de dos polinomios está definido salvo un factor constante. El resultado obtenido es relativamente complicado, como lo son los cálculos intermedios de la aplicación del algoritmo. Por ello, si se eligieran polinomios de grado mayor, los cálculos «explotarían».

No es de extrañar, pues, que los especialistas en cálculo formal hayan desplegado grandes esfuerzos a fin de encontrar un mejor algoritmo para determinar el *mcd*. Una de las soluciones propuestas es un algoritmo muy sofisticado en cuya prueba interviene el cálculo de matrices, creado en 1971 en Estados Unidos por W.S. Brown y G.E. Collins. Entre otras cualidades, este algoritmo da, para el caso anterior, un *mcd* de 260708, mucho más corto que el obtenido antes (6 cifras en vez de 19). Aunque muy sutiles, estos algoritmos para calcular el *mcd* sólo utilizan las matemáticas que se enseñan en los primeros años de universidad. He aquí un problema más arduo: determinar las primitivas de las funciones algebraicas. Recordemos que la primitiva de una función $f(x)$ es una función $g(x)$ tal que su derivada $g'(x)$ es igual a $f(x)$. Se trata, pues, de la noción inversa de la derivada. Se escribe $g(x) = \int f(x)$. Conocer la primitiva de una función puede servir para calcular la integral de dicha función en un intervalo.

Primitiva elemental. La primitiva de un polinomio es un polinomio, pero la primitiva de una fracción racional en general no es una fracción racional: se necesitan logaritmos para expresarla. Con mayor generalidad, llamemos «elemental» a una función que se expresa en términos de funciones usuales (radicales, logaritmos, exponenciales y funciones trigonométricas). ¿Es elemental la primitiva de una función elemental? La respuesta es no. Consideremos la función elemental $f(x) = \exp(x^2)$; esta función admite ciertamente una primitiva, aunque no una primitiva elemental. Éste es el caso general: raras son las funciones que admiten una primitiva elemental.

Al buscar la primitiva de una función se plantea, pues, un problema: ¿se expresa dicha primitiva por medio de funciones elementales y, en caso afirmativo, de qué modo? Por supuesto, se busca también un algoritmo eficaz para responder a esta cuestión. El problema ha sido estudiado desde el siglo XIX. La aparición de programas de cálculo formal le ha dado un nuevo aliento, pues los cálculos son demasiado complejos para ejecutarse sin ordenador.

Un caso particular importante es el de las llamadas funciones algebraicas, que son funciones de la forma $P(x, y)/Q(x, y)$, donde *P* y *Q* son polinomios y además *y* depende algebraicamente de *x*, es decir, $R(x, y) = 0$ para un cierto polinomio R. Así, la expresión: $(y^2 + xy + 1)/(y + 3x)$, con $y^4 + x^2y + x^3 = 0$, constituye una función algebraica. Un paso fundamental fue el dado por Joseph Liouville en 1833. Este matemático francés demostró que, caso de existir, una primitiva elemental debe tener una forma muy simple (en la que tal vez intervienen logaritmos, pero no exponenciales ni funciones trigonométricas, ni siquiera logaritmos de logaritmos).



Joseph Liouville, gran matemático francés del siglo pasado, demostró un teorema clave relativo a las primitivas de las funciones. Cien años después, el cálculo de primitivas sigue siendo objeto de intensas investigaciones. (Foto J.-L. Charmet)

En 1969, el estadounidense R.H. Risch dedujo de lo anterior un método que responde al problema de las primitivas de las funciones algebraicas. Este método se basa en un «cálculo» aplicado a los puntos de la curva plana de ecuación $R(x, y) = 0$. Pero no se sabía cómo automatizar este cálculo, por lo que no se trataba todavía de un algoritmo. Sólo a raíz de los trabajos del británico James Davenport en 1981 y del norteamericano Barry Trager en 1984 se consiguió implantar el método de Risch.

Se ha llegado todavía más lejos en el caso particular de las funciones hiperelípticas, unas funciones algebraicas de la forma:

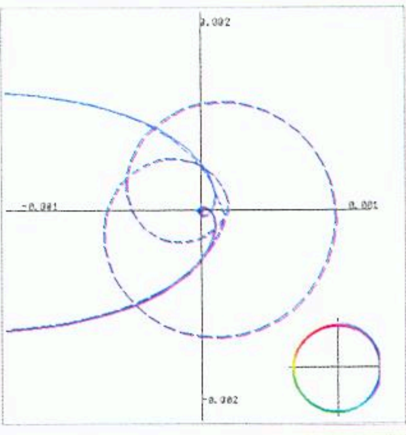
$f(x) = P(x,\sqrt{r(x)})/Q(x, \sqrt{r(x)})$, donde $r(x)$ es un polinomio. Estas funciones corresponden a $R(x, y) = y^2 - r(x)$. Los científicos topan a menudo con ellas en la práctica. Un ejemplo de una tal función es:
$f(x) = (21x^5 + 5x^4 - 3x^3 + 12x^2 + 2x + 2)/(2xy)$ donde $y = \sqrt{(x^3 + 1)}$ (aquí se cumple, por tanto, $r(x) = x^3 + 1$ y $R(x, y) = y^2 - (x^3 + 1)$).
Esta función admite una primitiva elemental, a saber:
$\int f(x) = (3x^2 + x - 1)y + 1/3 \log((y-1)^3 (y+1)/x^6)$.

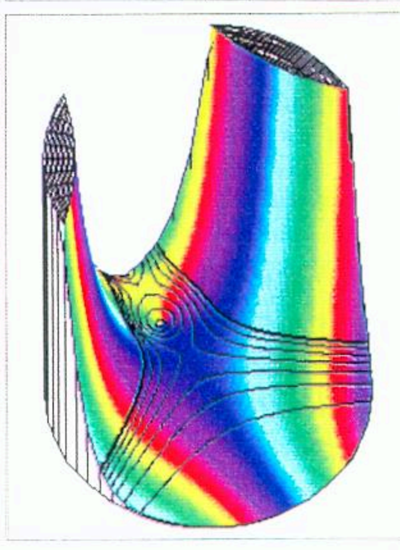
En cinco segundos. En 1994, Laurent Bertrand, de la Universidad de Limoges, elaboró un algoritmo reservado a las funciones hiperelípticas mucho más eficaz que los precedentes porque representa mejor la curva $R(x, y) = 0$. En el caso de la función anterior, por ejemplo, se obtiene el resultado diez veces más deprisa que antes (unos cinco segundos en vez de los cincuenta de Maple con una máquina IBM RS 6000).
Por otra parte, un investigador francés que trabaja en IBM, en Estados Unidos, Manuel Bronstein, consiguió en el año 1990 aplicar el método de Risch a ciertos tipos de funciones no algebraicas. A la vista de los progresos realizados, se puede apostar sin riesgo por próximas mejoras y generalizaciones de los resultados.
Los programas de cálculo formal no siempre son fáciles de utilizar. En general, hay que guiar los cálculos. La elección de la representación, del algoritmo y de los métodos de simplificación es fundamental. Lamentablemente, la que realiza automáticamente el programa no siempre es la que mejor se adapta al problema particular planteado.

Un tiempo de cálculo exponencial. Supongamos, por ejemplo, que deseamos conocer las dos últimas cifras del número 19969^{19969}. Se puede pedir al programa que calcule dicho número y luego lo divida por 100. El resto de la división está formado por las dos cifras buscadas, a saber, 29. Pero algunos conocimientos básicos de aritmética demuestran que es posible sustituir cada uno de los resultados intermedios del cálculo por el resto de su división por 100. Se manejan así números mucho más pequeños, con lo que se gana un tiempo apreciable (un factor 1000 en nuestro ejemplo).

Ecuaciones diferenciales: el programa DESIR

DESIR (*Differential Equations, Singularities Irregular and Regular*) es un ejemplo de programa que integra métodos formales, numéricos y gráficos. Este software, creado en los años 1980 por los equipos de Jean Della Dora y Jean-Pierre Ramis, de las Universidades de Grenoble y Estrasburgo, se ocupa de la resolución de ecuaciones diferenciales lineales en el ámbito de los números complejos. DESIR recurre a sofisticadas técnicas formales para obtener, tras una etapa numérica, una excelente aproximación de las soluciones cerca de las singularidades (puntos donde ciertas cantidades se hacen infinitas), precisamente allí donde los métodos numéricos clásicos son inutilizables. Luego, estos valores aproximados son representados gráficamente por un programa adaptado. En efecto, estas funciones, que a todo número complejo le asocian otro número complejo, no pueden dibujarse directamente. Como un número complejo equivale a un punto del plano (espacio bidimensional), la gráfica requeriría un espacio de cuatro dimensiones. DESIR propone dos métodos de visualización. Los dibujos reproducidos aquí representan la llamada función de Airy, solución particular de la ecuación $f''(z) - zf(z) = 0$.

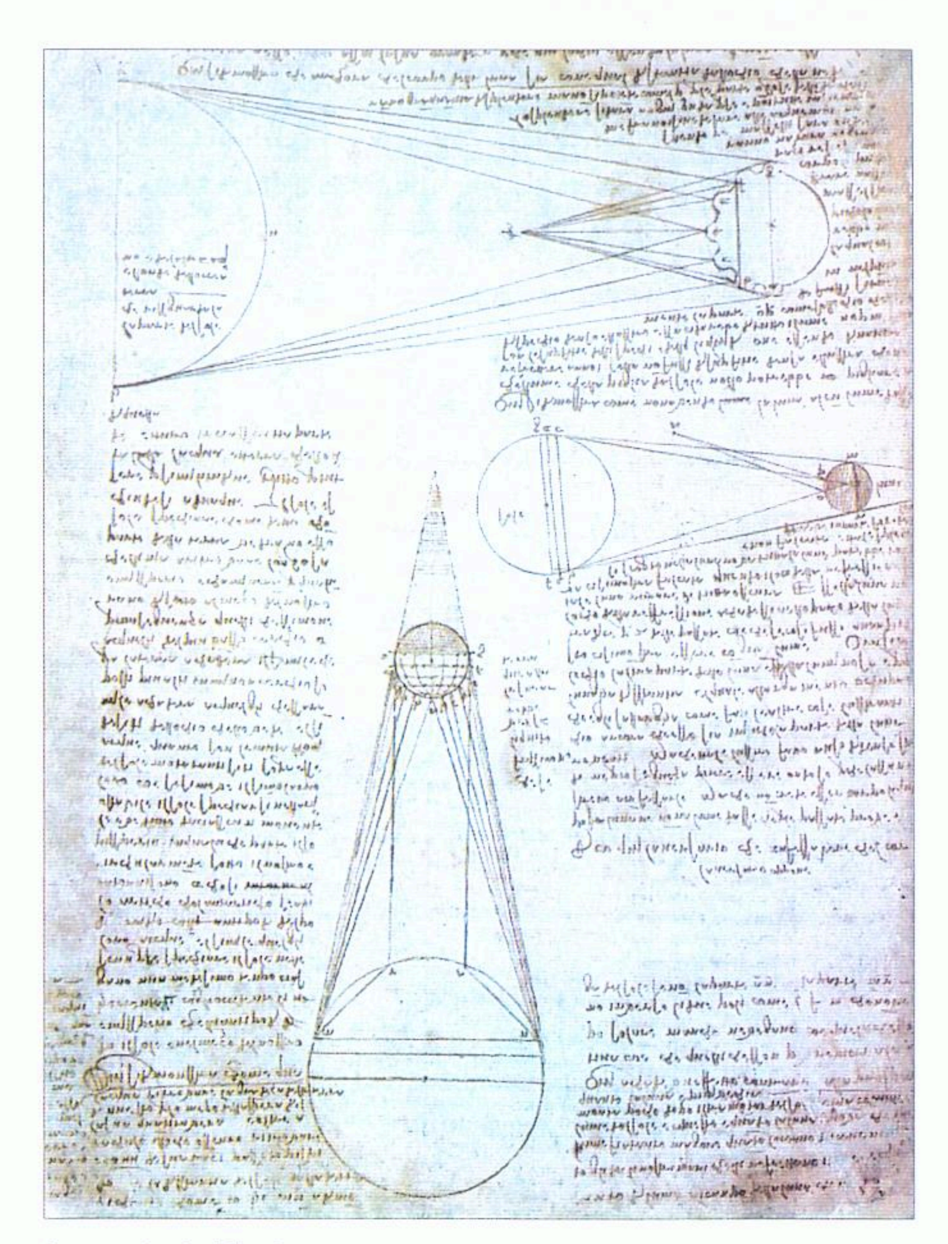




El primer método (A) describe la imagen por la función *f* de una curva, concretamente una circunferencia de centro 0 y radio 6. La imagen de esta circunferencia es una curva del plano. Se conviene en colorear de la misma manera un punto z de la circunferencia de partida y su imagen *f*(z) de la curva. A la escala inicial, la imagen de la parte derecha de la circunferencia es invisible: de hecho está concentrada cerca del origen, como cabe constatar pidiendo al programa que lleve a cabo «zooms» (B). Un segundo método de visualización (C) consiste en aplicar la función *f* a los puntos de una superficie simple, aquí el disco de centro 0 y radio 4. En esta representación gráfica, a cada punto z del disco se le atribuye una altura igual al módulo de *f*(z) y un color que codifica el argumento de *f*(z). Aquí, la superficie obtenida ha sido truncada a la altura 10 y se han representado algunas líneas de nivel. (Documents Laboratoire Modélisation et calcul/INPG-Grenoble)

Con mayor generalidad, se pueden encontrar tres tipos de escollos en la utilización de un sistema de cálculo formal. El primero estriba en que un método válido para un problema de pequeño tamaño no suele serlo para un problema análogo de mayor tamaño. En efecto, los algoritmos suelen tener una gran complejidad, en el sentido de que el tiempo de cálculo crece muy deprisa con el tamaño de los datos de entrada: los tiempos de cálculo

exponenciales —tiempos que se elevan al cuadrado cuando el tamaño de los datos se dobla— son frecuentes, los doblemente exponenciales —exponencial de una exponencial— no son infrecuentes. Recordemos que los cálculos formales proceden de manera exacta, con enteros «arbitrariamente largos». Por consiguiente, al incrementarse el tamaño de los datos no sólo aumenta el número de las operaciones sino también el tamaño de los coeficientes sobre los cuales se efectúan dichas operaciones (véase el recuadro: «Robótica y sistemas polinómicos: el programa GB»). Por ello, gran parte de las investigaciones de cálculo formal están dedicadas al estudio de la complejidad de los algoritmos. ¿Cuánto tiempo y cuánto espacio de memoria son necesarios para ejecutarlos, en función del tamaño de los datos de entrada? También es importante, aunque generalmente todavía más difícil, averiguar la complejidad «intrínseca» de un problema. Durante mucho tiempo se creyó, por ejemplo, que la factorización de los polinomios de coeficientes enteros era un problema de complejidad exponencial. Ahora bien, en 1982 se descubrió un algoritmo polinómico* gracias a los trabajos de los holandeses A.K. Lenstra y H.W. Lenstra Jr. y el húngaro L. Lovacz. No obstante, los sistemas de cálculo formal siguen uitilizando los antiguos métodos de factorización: sólo para polinomios de grado muy elevado (del orden de varios cientos), que actualmente nadie sabe resolver, podría resultar más eficaz el nuevo algoritmo. Este ejemplo pone de manifiesto que al elegir e implantar algoritmos los estudios teóricos sobre complejidad constituyen una guía extremadamente útil, aunque deben completarse con una experimentación cuidadosa.



Leonardo da Vinci
La distancia entre el Sol y la Tierra y la medida de la Luna. Cod. Leicester, fol. 1v, Royal Library, Windsor Castle.

Soluciones complejas. Una segunda dificultad estriba en la manera de plantear el problema. Sabemos más o menos qué significa «resolver» un sistema de ecuaciones lineales. ¿Sabemos también qué es «resolver» un sistema de ecuaciones polinómicas o diferenciales? Algunos usuarios quieren saber simplemente si hay alguna solución, otros si el número de soluciones es finito, etc. Pero incluso cuando el número de soluciones es finito se impone alguna elección. Por ejemplo, ¿qué significa resolver la ecuación (E): $x^4 - 2x^3 + x - 1 = 0$? Si nos interesamos sólo por las soluciones pertenecientes al conjunto de los números complejos* resulta que hay cuatro de tales soluciones. He aquí varias respuestas «razonables»:

- Un valor aproximado de las cuatro soluciones ($i = \sqrt{-1}$):
1,866760400; –0,8667604000;
0,5000000000 ± 0,6066580495i

Estos valores, aunque satisfactorios como resultados del cálculo, impiden seguir realizando cálculos exactos: llamando a = 1,866760400 al valor aproximado de la primera solución, entonces $a^4 - 2a^3 + a$ no vale 1 sino 1,000000010.

-una expresión por radicales de las soluciones:
$1/2 \pm 1/2\sqrt{(3 + 2\sqrt{5})}$ y $1/2 \pm 1/2\sqrt{(3 - 2\sqrt{5})}$.

Se trata de un resultado exacto, con el cual es posible seguir realizando cálculos exactos. Pero es bastante complicado y no es generalizable. Así, las soluciones de grado ≥ 5 (como $x^5 - x + 1 = 0$) no suelen poder expresarse con radicales.

-Una regla de cálculo: las soluciones de la ecuación (E) son los números a tales que:
$a4 = 2a^3 - a + 1$.

Esto permite simplificar todas las expresiones que dependen de una solución a sustituyendo a^4 por $2a^3 - a + 1$, a^5 por $4a^3 - a^2 - a + 2$, etc. Por otro lado, el símbolo a representa una raíz cualquiera de la ecuación (E), por lo que engloba las cuatro soluciones. Pese a su aparente ingenuidad, este método es con frecuencia el mejor para los cálculos intermedios.

Algunos ejemplos de las posibilidades de un sistema de cálculo formal, dados aquí por el programa Maple

•desarrollar una expresión simbólica:
desarrollo ((a+b) * (a-b));
$a^2 - b^2$

• *desarrollar (a + b)(a −b)*

•factorizar un polinomio:
factor (x^3-6*x^2+11*x-6) ;
(x - 1) (x - 2) (x - 3)

• *factorizar el polinomio $x^3 - 6x^2 + 11x - 6$*

•factorizar un número entero :
factor 30 ! ;
$(2)^{26} (3)^{14} (5)^{7} (7)^{4} (11)^{2} (13)^{2} (17) (19) (23) (29)$

• *descomponer 30 ! en factores primos*

•calcular una integral :
integral (1/x^2, x=1..infinito) ;
1

• *calcular la integral : $\int_1^\infty 1/x^2\,dx$*

•calcular un límite :
límite (sen (x) / x, x=0) ;
1

• *calcular el límite $\lim_{x \to 0} (\text{sen } x)/x$*

•calcular un desarrollo en serie :
series (sqrt (cos (x)) , x=0 , 8) ;
$1 - \frac{1}{4}x^2 - \frac{1}{96}x^4 - \frac{19}{5760}x^6 + o(x^8)$

• *calcular el desarrollo en serie de √(cos x) en las proximidades de x = 0 hasta el octavo orden 8*

perguntas — respuestas

Figura 1. La figura muestra los datos presentados por el ordenador durante una sesión de demostración con el programa Maple (en rojo, las preguntas formuladas por el usuario; de azul, las respuestas calculadas por el programa; los comentarios están en negro).

Tercera dificultad: la representación de los resultados, que depende mucho del programa considerado. Por ejemplo, para el sistema Maple, la instrucción: p: = $(x - 1)$* $(x - 2)$ ^2 * $(x-3)$ ^ 3 va seguida de la respuesta $(x - 1)(x - 2)^2(x - 3)^3$. Hay que recurrir a una función de simplificación para poder obtener la forma desarrollada ($x^6 - 14x^5 + 80x^4 - 238x^3 + 387x^2 - 324x + 108$). En el sistema Axiom, en cambio, es la forma desarrollada la que es presentada por defecto, por lo que es necesaria una instrucción para pasar a la escritura factorizada.

Todos los casos posibles. A veces, el resultado es tan largo que no es posible utilizarlo directamente. El programa GB (véase recuadro) puede suministrar un resultado que ocupa cientos de megaoctetos; «imprimirlo» no resuelve gran cosa. Solamente un tratamiento ulterior del resultado puede aportar informaciones interesantes sobre las soluciones. Antes de terminar, volvamos al manejo de los símbolos. Sabemos que se trata de una especificidad de los sistemas de cálculo formal. Estos últimos son capaces de resolver la ecuación $ax = b$ (resultado $x = b/a$) sin saber nada ni de a ni de b y evaluar la expresión $(1 + x - 1)/x$(resultado: 1) sin saber el valor de x. Estos resultados son correctos, pero no siempre: no se tienen en cuenta los casos particulares. El sistema, por ejemplo, no «ve» que la ecuación $ax = b$ tiene infinitas soluciones cuando $a = b = 0$ y no tienen ninguna cuando $a = 0$ y $b \neq 0$. Tampoco se da cuenta de que la expresión $(1 + x - 1)/x$ no está definida cuando $x = 0$. Distintos trabajos en curso pretenden tratar automáticamente todos los casos posibles en los cálculos simbólicos. Un método general es la «evaluación dinámica», basada en la automatización del método ingenuo: a medida que el algoritmo se va ejecutando se toman en cuenta todas las posibilidades y los distintos cálculos subsiguientes se realizan en paralelo. Este método ha sido implantado en Axiom por la matemática española Teresa Gómez Díaz, de la Universidad de Limoges, para algunos tipos de problemas. Pero todavía queda mucho que hacer en esta dirección.

Un gran iceberg. Todas estas dificultades, y otras de las que no se ha podido hablar aquí, son objeto de activas investigaciones a caballo entre las matemáticas y la informática, con progresos debidos a un vaivén entre la teoría y la práctica. Sin duda alguna, el cálculo formal constituye un maravilloso instrumento con el cual los científicos no se atrevían a soñar siquiera hace unos decenios. Para evitar las desilusiones, sin embargo, hay que recordar que un sistema de cálculo formal esconde bajo una apariencia simple unos instrumentos muy potentes y por tanto difíciles de manejar.

No hay que creer, por último, que el cálculo formal sea sólo un instrumento. Aunque su objetivo básico consiste, en efecto, en elaborar unos procedimientos automáticos de cálculo simbólico, el camino hacia ellos recorre varios campos de la informática y las matemáticas. Si bien en este artículo no ha quedado del todo patente, el cálculo formal constituye para estas disciplinas una fuente de nuevas problemáticas de investigación. Lo que ve un usuario en su pantalla de ordenador no es sino la parte emergida de un gran iceberg.

DOMINIQUE DUVAL es profesora de la Universidad de Limoges y responsable del Laboratorio de aritmética, cálculo formal y optimización.

Mundo Científico ha publicado:
(I) Dominique Duval, «¿Cómo tratar los cálculos imposibles?», diciembre, nº 97, 1989.

*La DERIVADA de una función f en un punto x es el límite, cuando h tiende a 0, de cociente $(f(x+h)-f(x))/h$.

*En el caso de una sola variable x, UN POLINOMIO es una expresión de la forma $a_0 + a_1x + a_2x^2 + ... + a_nx^n$. El entero n es el grado del polinomio y los números a_i sus coeficientes. Una expresión como $1 + 6x + x^3y^2$ es un polinomio de dos variables, de grado 5 (= 3 + 2).

* UNA FRACCIÓN RACIONAL es una fracción $P(x)Q(x)$ en la que el numerador $P(x)$ y denominador $Q(x)$ son polinomios en la variable x.

* ALGORITMO POLINÓMICO
Para un tal algoritmo, el tiempo de cálculo T no aumenta más deprisa que un polinomio en n, el tamaño de los datos. Es un crecimiento mucho más lento que el de un «algoritmo exponencial», en el que T es una función exponencial de n.

*NÚMERO COMPLEJO
Número de la forma $z = x + iy$, donde x e y son números reales e $i = \sqrt{-1}$.

Para más información:

D. Lazard *et al.*, *Bulletin de liaison de la recherche en informatique et en automatique*, nº 130, 1991.
K.O. Geddes *et al.*, *Algorithms for Computer Algebra*, Kluwer Academic Publishers, 1992.
J. Davenport *et al.*, *Calcul formel. Systèmes et algorithmes de manipulations algébriques*, Masson, 2ª edición, 1993.
A.M. Cohen (ed.), *Computer Algebra in Indutry Wiley*, 1993.
D. Duval, «Algebraic numbers: an example of dynamic evaluation», *Journal of Symbolic Computation*, 18, 429, 1994.
Ian Stewart, *De aquí al infinito: las matemáticas de hoy*, Editorial Crítica, Barcelona, 1998.

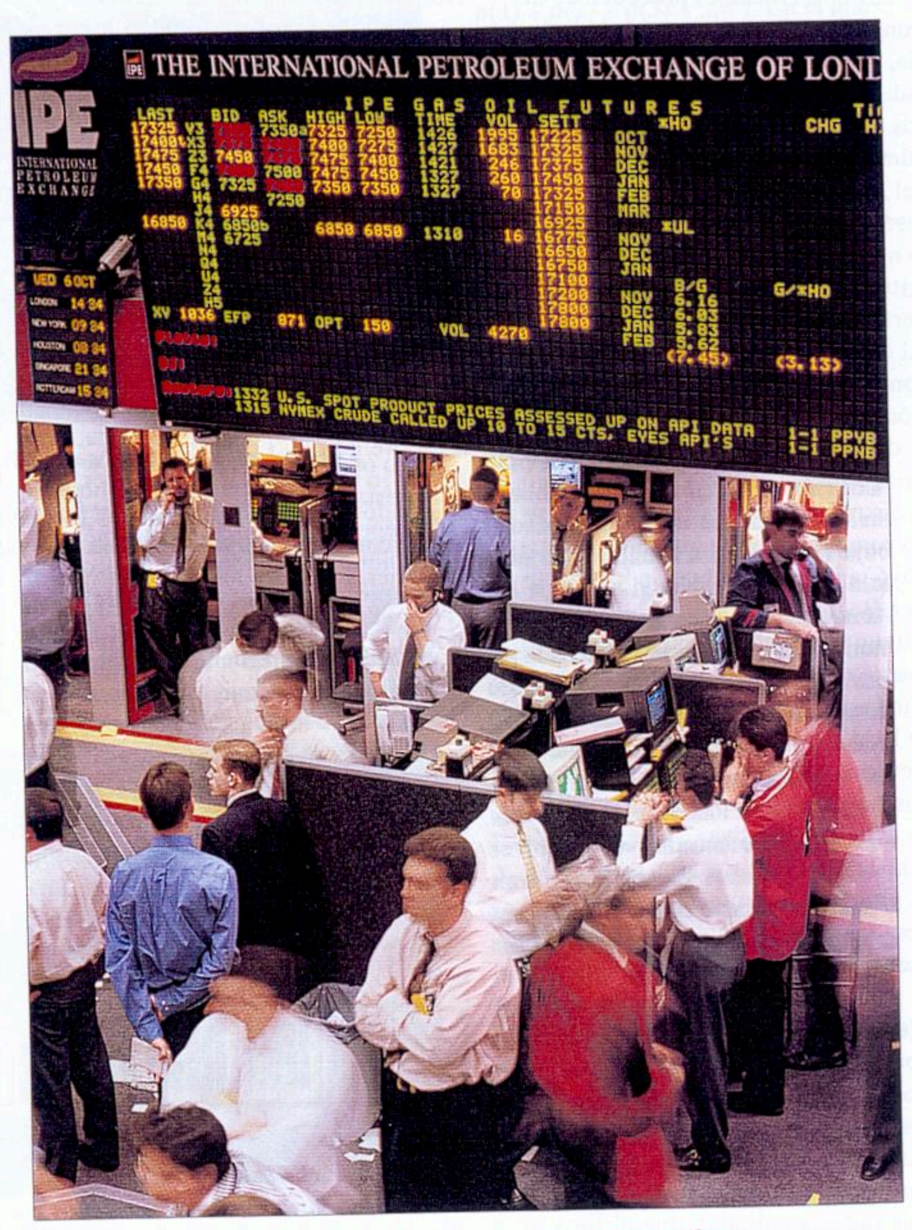
La primera cifra significativa

Ted Hill

¿Aparecen las cifras del 1 al 9 con la misma probabilidad? Según las medidas de frecuencias, no. Observado hace un siglo, este extraño fenómeno repercute en dominios tan alejados entre sí como las previsiones bursátiles, las constantes de la física y los censos demográficos.

Las nueve cifras, del 1 al 9, ¿aparecen con frecuencias diferentes en los titulares de los periódicos y en las cotizaciones de la Bolsa? La mayoría de las personas responderían sin vacilar que no. Sin embargo, en 1881, un artículo de dos páginas publicado en la revista *American Journal of Mathematics* y firmado por el astrónomo y matemático Simon Newcomb presentaba un fenómeno insólito: los primeros volúmenes de las tablas de logaritmos que equipaban ya toda biblioteca científica que se preciara eran más utilizados que los demás. Newcomb llegó a la conclusión de que los estudiantes y los investigadores trabajaban más a menudo con con números que empezaban por 1 que con números que empezaban por 2; estos últimos, a su vez, eran más utilizados que los que empezaban por 3 y así sucesivamente. Un razonamiento rápido y heurístico permitió a Newcomb llegar a la audaz conclusión de que la probabilidad de que un número elegido al azar dentro de un conjunto cualquiera tenga como primera cifra significativa d (siendo d un entero no nulo) viene dada por la fórmula siguiente: p («primera cifra significativa es igual a d») = $\log_{10}(1 + 1/d)$, con d = 1,2,...9.

Medio siglo de olvido. Ello suponía que la cifra 1 aparecía como primera cifra significativa aproximadamente un tercio de las veces. En cambio, la probabilidad de que la primera cifra significativa fuera un 9 era sólo del 4,6% (fig. 1).



No, las cifras significativas de los precios del petróleo no aparecen de una manera totalmente aleatoria: el 1 aparece mucho más a menudo que el 2, el 3, el 4, el 5, el 6, el 7, el 8 y el 9.¿Es este fenómeno curioso una característica inherente a nuestro modo de contar?(Foto J. Yeats/SPL/Cosmos)

El artículo de Newcomb pasó totalmente inadvertido. Cincuenta y siete años después, Franck Benford, físico de la General Electrics, hacía la misma observación, basada también en tablas de logaritmos.

Benford llegó a la misma ley de probabilidad logarítmica.

Más tarde, Benford verificó su conjetura por medio de gran número de datos. Según dijo él mismo, su investigación fue *«tan amplia como se lo permitieron el tiempo y la energía humanamente disponible»*. Benford estuvo varios años reuniendo y tabulando datos, unos datos que publicó en 1938 en un artículo de los *Proceedings of the American Philosophical Society*.(1) Este documento contiene 20 229 observaciones procedentes de campos tan dispares como la hidrología, las estadísticas de la liga estadounidense de béisbol, los pesos atómicos de los elementos químicos y números extraídos al azar de artículos del *Reader's Digest*. Su tabla de primeras cifras significativas —situadas lo más a la izquierda posible en la escritura de los números— concuerda perfectamente con su ley.

Al contrario que el artículo de Newcomb, el de Benford llamó la atención, tal vez en parte por tener la suerte de aparecer junto a un artículo de física también llamado al éxito. Como la distribución de Newcomb había sido completamente silenciada, se atribuyó la ley de distribución logarítmica a Benford.

Aunque esta última ha sido ampliamente reconocida y aplicada en la segunda mitad del siglo XX, todavía no se ha dado de ella una demostración rigurosa. ¿Qué principios matemáticos subtienden esta «ley de la primera cifra significativa»? Los primeros matemáticos que trataron de resolver el problema no pudieron hacer gran cosa más que inventariar exhaustivamente las tablas de datos conformes.

Ejemplos y contraejemplos. Estos conjuntos son numerosos, pero de todos modos hay contraejemplos. Así, los números de teléfono de una misma región suelen empezar por la misma cifra, lo cual falsea la conjetura. Las raíces cuadradas, que al fin y al cabo constituyen una tabla neutra desde este punto de vista, tampoco siguen esta ley. Lo cual no es óbice para que desde el artículo de Benford los «buenos» ejemplos de estas colecciones numerales hayan empezado a proliferar de manera sorprendente. Las constantes físicas, los números que aparecen en los titulares de los periódicos, los datos de libros de contabilidad o de registros de cálculos científicos concuerdan también con la distribución logarítmica.

La ley logarítmica predice que la segunda ley significativa se distribuye mucho más uniformemente que la primera

Donald Knuth, informático de la Universidad de Stanford, dedicó a sus aplicaciones todo un capítulo de su obra clásica de 1968, *The Art of Computing Programming*. Más recientemente, en 1996, Eduardo Ley, analista del Instituto de recursos del futuro, del distrito de Washington, demostró que los indicadores bursátiles como el Dow Jones de la cotización media de los valores industriales, o también el del nivel de pobreza, siguen de cerca la ley de Benford. Mark Nigrini, profesor de contabilidad, autor de una tesis sobre las aplicaciones de esta ley, hizo otro tanto con las cifras del censo estadounidense de 1990. El lector escéptico podrá sacar sus propias conclusiones anotando los números que aparecen en primera página de su periódico local, o, como sugiere Knuth, eligiendo aleatoriamente datos de un almanaque agrícola.

No obstante, cuando se considera en forma generalizada, la ley logarítmica tiene un alcance muy distinto. En efecto, en su formulación cabe incluir no sólo la segunda cifra significativa (que puede ser nula) sino también las demás —incluso las que están a la derecha de la coma—. En esta forma general, la ley dirá, por ejemplo, que la probabilidad de que un número, en una serie de datos, tenga como primeras tres cifras significativas 3, 1 y 4 vale $\log_{10}(1 + 1/314) \approx 0{,}0014$. La segunda cifra, cuyo valor, en nuestro ejemplo, aparece en las decenas del denominador de la fracción, está distribuida mucho más uniformemente que la primera. En efecto, su influencia es menor en un orden de magnitud a la de la cifra de las centenas. Lo mismo ocurre con las demás cifras significativas. Al movernos hacia la derecha, las diferencias entre las probabilidades de las ocurrencias se hacen cada vez más tenues. La ley pierde su validez y da paso a la equiprobabilidad habitual. Se observa, por ejemplo, que los distintos valores posibles para la quinta cifra aparecen con unas probabilidades que son casi todas iguales a un décimo, una tendencia que se confirma para los órdenes siguientes.

Una demostración delicada. La ley general implica también que las cifras significativas no son independientes y que la presencia de una cifra modifica la probabilidad de aparición de las demás. Por ejemplo, un cálculo simple demuestra que la probabilidad de que la segunda cifra sea 2, sin imponer

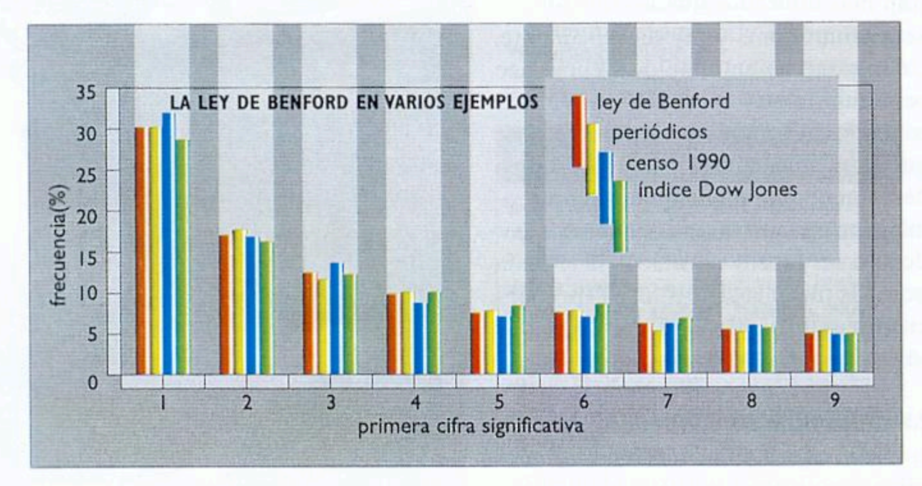
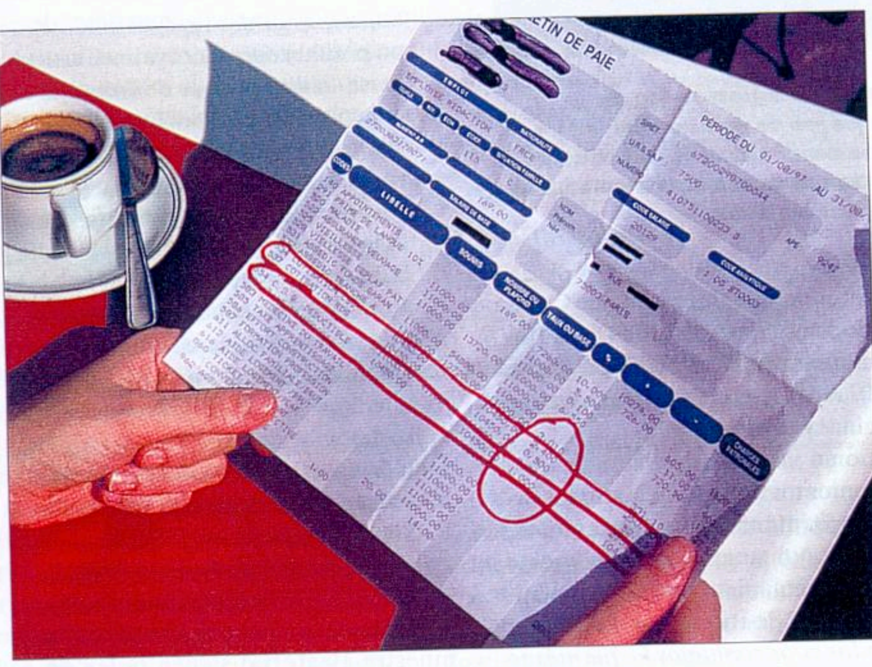


Figura 1. La ley de Benford predice frecuencias decrecientes para la aparición, como primera cifra significativa, de los enteros del 1 al 9. Varios conjuntos de datos la siguen naturalmente: los números tomados de los titulares de los periódicos, los del censo estadounidense de 1990 o los del índice Dow Jones.

ninguna condición a las demás, es del orden de 0,109, pero la de que dicha cifra sea 2 sabiendo que la primera es 1 es del orden de 0,115. En los años 1960, muchos aficionados interesados por el artículo de Benford que trataron de demostrar la ley toparon con dos escollos de peso. El primero es muy simple: sólo ciertos conjuntos de datos responden a la ley, pero no hay ningún criterio estadístico que permita predecir esta adecuación. Siguiendo un camino opuesto, los matemáticos trataron de demostrar que la ley logarítmica era una característica inherente a nuestro sistema numérico, explicando así la frecuencia de sus manifestaciones empíricas. Debían demostrar, pues, que todos los números la seguían, para lo cual emplearon con mayor o menor fortuna los instrumentos clásicos de la disciplina: cálculo de promedios, integración o experimentos aleatorios como sorteos de bolas. Una hipótesis comúnmente admitida consiste en suponer invariante la escala de datos. Esta idea corresponde a la intuición de que una ley universal de las cifras significativas, de existir, deber ser independiente de las unidades utilizadas: por ejemplo, la ley debe funcionar siempre independientemente de que se elija el sistema métrico o un sistema de unidades anglosajonas. De hecho, la observación empírica prueba que los conjuntos de datos que la siguen continúan haciéndolo después de convertirlos a otras unidades. Esta operación se efectúa simplemente multiplicando por un coeficiente constante. Así, si una lista de precios en euros sigue, como había observado Eduardo Ley, la ley de Benford, su conversión en dólares, yens o pesetas no modifica significativamente las frecuencias calculadas y ello aunque la primera cifra significativa difiera radicalmente de una moneda a otra.[1] Asimismo, la utilización de datos expresados en euros por unidad conduce a las mismas frecuencias de cifras que su conversión a números de unidades por euro. Por contra, cabe observar que si el cuadro de números no sigue lo bastante de cerca la ley de Benford, como ocurre en una lista uniformemente distribuida (fig. 2), el cambio de moneda o la conversión de unidades modificará considerablemente la frecuencia de aparición de las distintas cifras. Se descubrió, por tanto, que sólo las series de números que siguen fielmente las frecuencias de cifras previstas por la ley de Benford siguen haciéndolo tras un cambio de escala.



¿Se detectarán pronto los fraudes fiscales gracias a la ley de Benford? Los trabajos de M. Nigrini son categóricos: comparado con la distribución de las cifras de una declaración real, un defraudador escribe demasiados 6 y demasiado pocos 1. Unos tests simples permiten detectar estos datos atípicos. (Foto X. Roest/Gamma)

Un segundo escollo cerraba el paso de los matemáticos: ninguna de las pruebas propuestas era rigurosa desde el punto de vista de la teoría de probabilidades. El hecho de que Newcomb y Benford hubieran enunciado la cuestión en términos adecuados —«*¿Cuál es la probabilidad de que la primera cifra significativa de un número sea d?*»— no cambiaba un ápice el asunto. En particular, se observó muy pronto que no se respetaba un axioma fundamental.

Cambio de escala. Como un suceso cierto posee, por definición, una probabilidad igual a la unidad, si se elige al azar un número positivo y si $p(1)$ es la probabilidad de que este número sea 1, $p(23)$ la de que este número sea 23 y así sucesivamente, se verifica: $p(1) + p(2) + p(3) + ... = 1$. En el marco de la ley de Benford, el estatuto de la cifra 1 era lo que impedía respetar esta propiedad.

La ley de Benford puede integrarse legítimamente al cálculo de probabilidades

Consideremos las dos leyes físicas $F = m\gamma$ y $E = mc^2$. En estas dos expresiones intervienen constantes universales. En la primera ecuación la elección de unidades a uno y otro lado del signo de igualdad hace que esta constante sea igual a 1 y pase inadvertida. En cambio, la de la segunda ecuación es la velocidad de la luz, c. Consideremos ahora la lista «exhaustiva» de todos sus valores obtenidos multiplicándolas sucesivamente por todos los enteros. Si, en tal conjunto, se distinguen las constantes que sólo sirven para ajustar las medidas y por tanto toman como valor numérico la cifra 1, esta cifra aparece ciertamente con una probabilidad p no nula. Por ello, en este mismo conjunto, el cambio de escala que consiste en doblar los valores hace aparecer una constante 2 que tiene *por lo menos* esta misma probabilidad p no nula. Y lo que vale para el 2 vale también para cualquier otro factor. Entonces, la probabilidad enunciada más arriba deja de satisfacerse, pues $p(1) + p(2) + ... + p(i)$ tiende a infinito y no a 1 cuando i tiende a infinito.

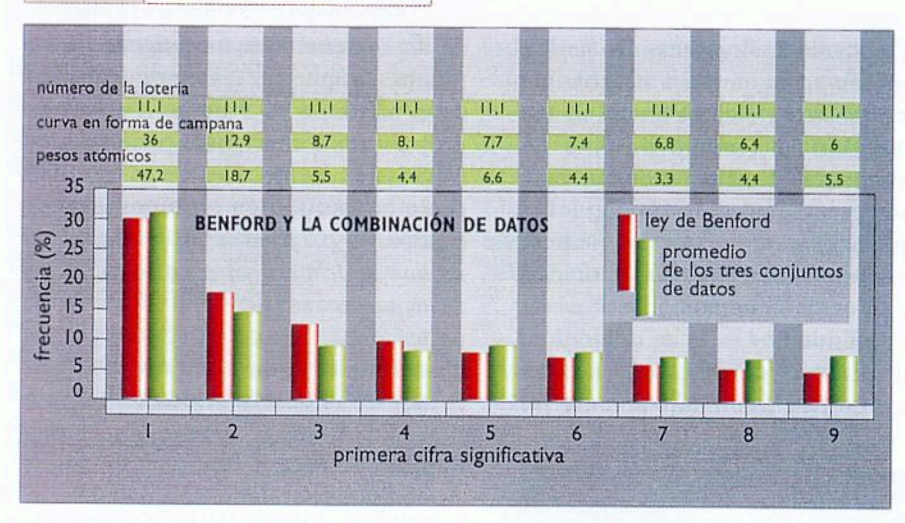


Figura 2. No todos los conjuntos de datos cumplen la ley de Benford: discrepan mucho de ella los de los juegos de azar (lotería), los de las curvas en forma de campana y los de los pesos atómicos. No obstante, el promedio de las tres series de frecuencias se acerca mucho a las predicciones de la ley, un fenómeno que ha llevado a reformular la misma.

Base métrica. Para soslayar el problema, supongamos que esta ley sea independiente de la base métrica, es decir, que sea igualmente válida en base 10, en base 100, en base 2 o en cualquier otra base —una suposición legítima cuando se intenta demostrar una ley universal—, válida para todos los números. Este supuesto implica la posibilidad de elegir una base de referencia en la que las constantes se expresen una sola vez. Así se pueden construir conjuntos equivalentes de constantes para una operación de cambio de base, lo que permite abarcar todos los números. Esta partición define unas «clases de equivalencia» que en términos probabilistas corresponden a sucesos idénticos, ignorados en el razonamiento precedente. La ley de Benford, por tanto, puede integrarse legítimamente al cálculo de probabilidades y los instrumentos de esta última se aplican a las ocurrencias de las cifras. Con estos instrumentos es posible demostrar rigurosamente que la ley logarítmica es también la única distribución de probabilidad invariante por cambio de escala (excluyendo la constante 1) e invariante por cambio de base.[2]

Reunión de conjuntos. Estos nuevos resultados, pese a haber quedado claramente establecidos en el sentido matemático, no ayudaban mucho a comprender intuitivamente las manifestaciones de la ley. ¿Qué tienen en común el censo estadístico de 1990, las tablas de logaritmos de 1880, las cifras que Benford tomó de los periódicos de 1930 y las cifras de los ordenadores de Knuth en los años 1960? Además, ¿por qué deberían seguir la ley logarítmica?

Como ya hemos observado, muchos conjuntos numéricos, incluidos los del propio Benford, «*no tienen esta forma*». No obstante, como advierte Ralph Raimi, matemático de la Universidad de Rochester, «*es la reunión de todos estos conjuntos lo que más se acerca a la ley de Benford*». Combínese la tabla de pesos moleculares con la de las estadísticas del béisbol y la de la superficie de los ríos franceses y se obtendrá un conjunto perfectamente ajustado a la ley (fig. 2).

Las muestras de varias distribuciones siguen la ley de Benford aunque no lo hagan las distintas distribuciones

En vez de pensar en una especie de tabla universal de todas las constantes posibles —el «conjunto de datos numéricos sacados de la librería mundial» de Raimi o los «*conjuntos imaginarios de números reales*» de Knuth, parece más natural seguir el ejemplo de Benford, quien consideraba que los datos podían proceder de varias distribuciones posibles y los buscaba en los periódicos o las cotizaciones bursátiles.

Desarrollando esta idea por medio de la teoría de probabilidades y de los recientes resultados sobre invariancia por cambio de escala y cambio de base, fue bastante fácil llegar, en 1996, a una nueva formulación estadística de la ley de la primera cifra significativa: si las distribuciones se eligen al azar (mediante cualquier método «no sesgado») y se extraen muestras aleatorias, las frecuencias de las cifras significativas siguen la ley de Benford aunque no lo hagan las distribuciones de partida.[3]

Supongamos que recogemos datos de un periódico y que el primer artículo versa sobre números de lotería, el segundo sobre una población particular distribuida siguiendo la célebre curva en forma de campana y la tercera sobre la última puesta al día de la tabla de pesos atómicos. Ninguna de estas distribuciones sigue la ley de Benford. No obstante, sí lo hace una muestra aleatoria que contenga valores de esos tres conjuntos. La puerta quedaba abierta a gran número de aplicaciones.

Una de ellas es el desarrollo matemático de los controles informáticos. Supongamos que se propone un nuevo modo de predicción de índices bursátiles (o de datos demográficos). Si los datos reales (por ejemplo, una muestra aleatoria) siguen la ley de Benford, es de esperar que el modelo prediga unos datos que sigan la misma ley, lo cual es fácil de verificar. No obstante, así utilizada, la ley no basta por sí sola para fundar un test de coherencia, pues no distingue entre los números 20 y 20 000 (estos números tienen 2 y 0 como cifras significativas).

Otra aplicación reciente tiene que ver con el diseño de los ordenadores. Es probable, en efecto, que los usuarios del futuro hagan trabajar sus ordenadores con números procedentes de varias distribuciones aleatorias no sesgadas. Los números utilizados por las futuras máquinas, por tanto, seguirán perfectamente la ley de Benford: no estarán distribuidos de manera totalmente uniforme, sino que obedecerán a una ley logarítmica. Una vez demostrada esta distribución particular, será posible sacar partido de unos ordenadores que desde un principio incorporen el conocimiento de los datos que tengan que manipular. Sabiendo que las cifras 9 son menos frecuentes que las 1

(y así sucesivamente en cualquier base), se podrá minimizar el volumen de almacenamiento de los datos u optimizar los flujos.

¿El fin del cálculo binario? La idea subyacente es simple: piénsese en una caja registradora en lugar de un ordenador. Conociendo la frecuencia de las transacciones y las distintas clases de artículos vendidos, se podría diseñar perfectamente una caja que tuviera en cuenta esta distribución numérica. Por ejemplo, podría contener cajones de tamaño variable según las cifras o también recurrir a una disposición inteligente de estas últimas.

El matemático alemán Peter Schatte ha determinado que los ordenadores optmizados para tener en cuenta la ley de Benford deberían abandonar el cálculo binario y optar sin reservas por la base ocho. Otros investigan ya las ventajas, en términos de velocidad, de los ordenadores completamente logarítmicos.

Uno de los desarrollos más clásicos de esta ley procede de la contabilidad y se aplica a la detección de fraudes (errores o falsificación de datos). Nigrini ha recogido gran número de pruebas empíricas que justifican este empleo y su conclusión es que en muchos casos de observación la frecuencia de la primera cifra significativa sigue la ley de Benford (fig. 3). Si el fraude es deliberado, los datos falsificados raramente siguen la ley.

Al parecer, como saben los psicólogos, los defraudadores son incapaces de inventar unos datos realmente aleatorios. Uno de mis ejemplos favoritos al respecto procede de mi propia docencia. El primer día de clase, que inaugura un semestre de introducción a la teoría de probabilidades, pido a los estudiantes que hagan en casa el ejercicio siguiente. Si el nombre de soltera de su madre empieza por una letra entre A y L, deben echar doscientas veces una moneda al aire y anotar el resultado. En caso contrario, deben tratar de recrear sin ayuda de la moneda una tal secuencia de caras y cruces. A la mañana siguiente recojo los resultados y ante la sorpresa general separo las verdaderas secuencias de las falsas con un 95% de éxito.

En los programas más utilizados por los servicios de detección de fraudes van a incorporarse tests basados en la ley de Benford

Fraudes y falsificaciones. Aunque la demostración rigurosa es difícil, lo que hago es aplicar simplemente la regla siguiente: en un conjunto de doscientas tiradas reales, las sucesiones de seis caras o de seis cruces consecutivas aparecen con gran probabilidad. Una persona que trate de imitar resultados aleatorios raramente escribirá series homólogas tan largas.

La tesis doctoral sobre contabilidad de Nigrini propone un uso similar de la ley de Benford. Admitiendo, como indican estas investigaciones, que los datos normales de contabilidad siguen esta ley, Nigrini sugiere que fuertes discrepancias en las frecuencias de las cifras son otras tantas firmas de fraudes o falsificaciones.[4]

Este investigador ha elaborado tests de adecuación que permiten medir la discrepancia respecto al modelo de Benford. El Wall Street Journal informa de que gracias a los tests de Nigrini el servicio de fraudes de Brooklyn, en Nueva York, ha logrado detectar irregularidades en las contabilidades de siete compañías neoyorkinas. Por lo visto, los defraudadores suelen poner demasiado pocos datos que empiezan por 1 y demasiados que empiezan por 6 (fig. 3). Después de estos primeros éxitos, los servicios de impuestos de muchos Estados norteamericanos han solicitado el asesoramiento de Nigrini, cuyos tests se están incorporando a los programas más utilizados para la detección de fraudes.

Claroscuro. Cuando hace más de un cuarto de siglo se publicó en *Scientific American* el artículo de R. Raimi sobre la ley de Benford, el fenómeno de la primera cifra significativa fue considerado como una curiosidad matemática sin aplicaciones prácticas ni fundamentos teóricos satisfactorios.[5]

Este matemático escribió al respecto: «Todas las explicaciones anteriores [de la ley de Benford] dejan mucho que desear en lo tocante a la finalidad de la misma».

Ralph Raimi concluía en la oscuridad de la cuestión, pero hoy, aunque el capítulo final de la historia está todavía por escribir, el fenómeno aparece con una claridad mucho mayor. Su descripción, basada en el cálculo de probabilidades, propone importantes aplicaciones para los problemas de nuestra sociedad.

BENFORD Y LOS FRAUDES

Figura 3. Los 5 y los 6 predominan netamente en los datos falseados, pero las series de seis cifras significativas aleatorias reconstruidas por 743 voluntarios muestran una distribución equitativa. Sólo las frecuencias de las cifras significativas de los datos fiscales auténticos obedecen a la ley de Benford.

TED HILL es profesor de matemáticas del Instituto de tecnología de Georgia, en Atlanta (Estados Unidos). Tomado de un artículo publicado por la revista *American Scientist* (nº 86, 1998).

(1) E.Ley, *American Statistician*, 50, 311, 1996.
(2) F. Benford, *Proceedings of the American Philosophical Society*, 78, 551, 1938.
(3) T. Hill, *Statistical Science*, 10, 354, 1996.
(4) M. Nigrini, *Journal of the American Taxation Association*, 18, 72, 1996.
(5) R. Raimi, *Scientific American*, 109, 119, 1969.

Ordenadores en busca de aritmética

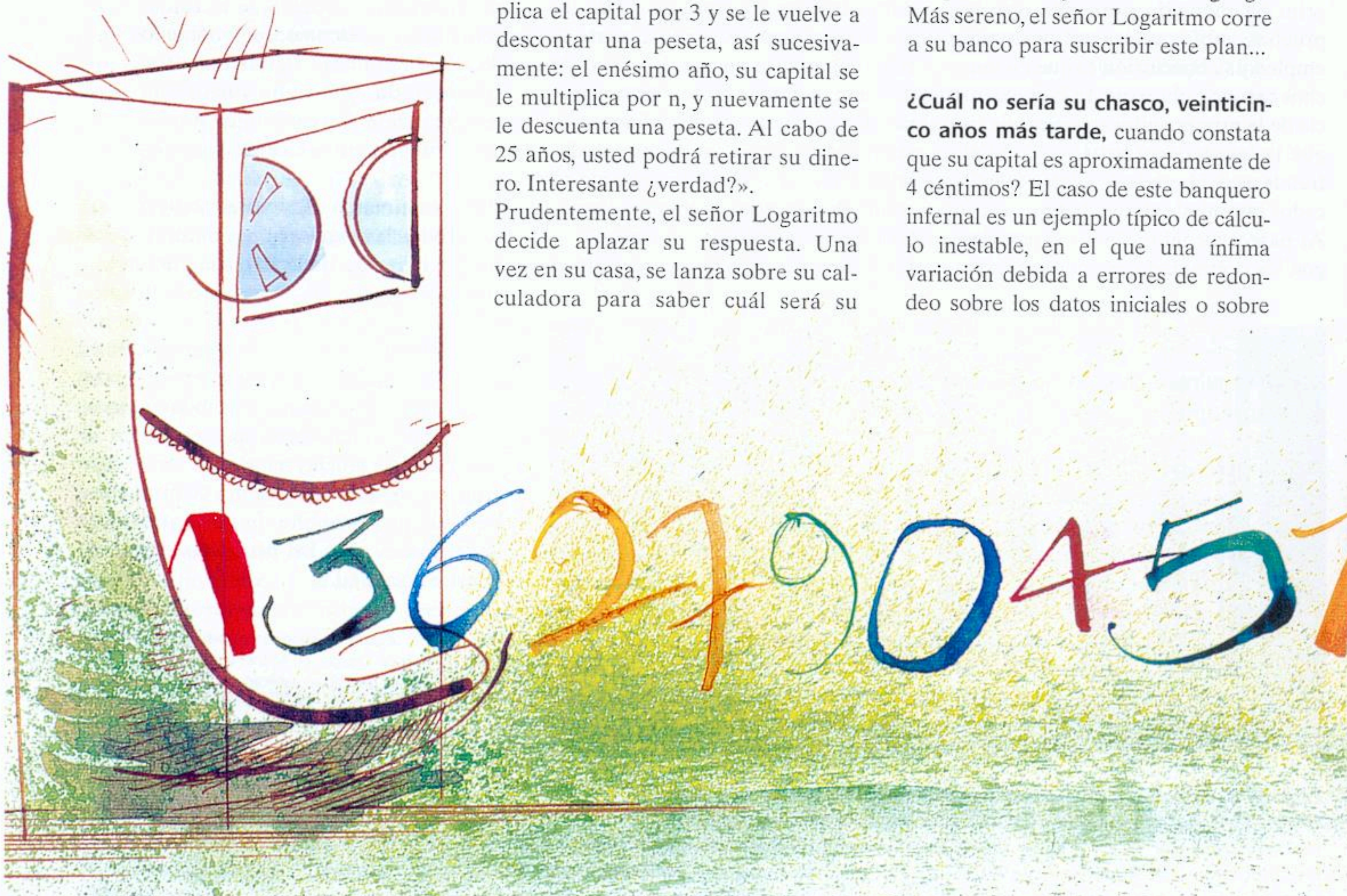
Jean-Michel Muller

Debido a errores de redondeo, un ordenador potente es susceptible de dar un resultado completamente falso en cálculos en «coma flotante». Las consecuencias prácticas son potencialmente devastadoras. Hoy se está investigando activamente sistemas de representación de los números más adecuados para los ordenadores.

Alfredo Logaritmo desea hacer una imposición bancaria a largo plazo con el fin de asegurar el porvenir de su descendencia. Se informa cerca de la Sociedad Caótica de Banca sobre su nuevo plan de ahorro. El banquero le da esta explicación: «Usted hace una aportación inicial de (e–1)pesetas (e=2,718281828459045... es la base de los logaritmos naturales). El primer año, usted pierde: su capital se le multiplica por 1 y se le descuenta una peseta por gastos de gestión. El segundo año irá mejor: su capital se le multiplica por dos y se le sigue descontando una peseta por gastos de gestión. El tercer año, se le multiplica el capital por 3 y se le vuelve a descontar una peseta, así sucesivamente: el enésimo año, su capital se le multiplica por n, y nuevamente se le descuenta una peseta. Al cabo de 25 años, usted podrá retirar su dinero. Interesante ¿verdad?».

Prudentemente, el señor Logaritmo decide aplazar su respuesta. Una vez en su casa, se lanza sobre su calculadora para saber cuál será su capital dentro de veinticinco años. El resultado le asusta: ¡será menos de 140.302.545.600.000 pesetas! ¿Es que este nuevo plan de ahorro resultará ser una estafa? Para asegurarse, pide a un amigo informático que efectúe el cálculo en una máquina más potente: obtiene entonces un resultado de +4.645.987.753 pesetas, lo que constituye una buena ganga. Más sereno, el señor Logaritmo corre a su banco para suscribir este plan...

¿Cuál no sería su chasco, veinticinco años más tarde, cuando constata que su capital es aproximadamente de 4 céntimos? El caso de este banquero infernal es un ejemplo típico de cálculo inestable, en el que una ínfima variación debida a errores de redondeo sobre los datos iniciales o sobre



un resultado intermedio puede provocar una variación enorme sobre el resultado final. La inestabilidad es tal que ninguna máquina, por precisa que sea (la calculadora utilizada en el caso anterior es una CASIO FX702P, y el ordenador un SUN SPARC, dos máquinas de excelente calidad numérica), no puede calcular el resultado correcto, a pesar de que, *a priori*, este cálculo parece muy simple: se efectúan 25 multiplicaciones por pequeños enteros y se resta 25 veces el número 1.

De esta anécdota se deduce que los usuarios se forman muchas veces ideas falsas sobre los ordenadores. O bien les atribuyen una fiabilidad superior a la que tienen en realidad, o bien piensan que los algoritmos de operación de los ordenadores no pueden ser más que variantes de los que hemos aprendido en la escuela. Empezaré, pues, por presentar algunos problemas relacionados con la manipulación de los números, antes de mostrar cómo pueden ponerse a punto algoritmos muy diferentes de los clásicos.

En el ejemplo descrito anteriormente, se obtenían resultados visiblemente aberrantes, lo que podía despertar sospechas. Pero no siempre es así y, a veces, se observa en máquina una convergencia buena y rápida hacia un resultado totalmente falso, tal como demuestra este segundo ejemplo. En efecto, consideremos la serie (a_n) de números, empezando por $a_0 = 2$, $a_1 = -4$ y cuyos términos siguientes vienen definidos por la relación : $a_{n+1} = 111 - 1\,130/a_n + 3000/\,a_n\,a_{n-1}$. El límite de a_n es 6 (se dice que una serie converge hacia un límite λ, o tiene un límite λ, si sus términos se aproximan cada vez más a λ, de tal manera que, para un umbral determinado, por pequeño que sea, todos los términos de la serie se encuentran, a partir de un cierto rango, a una distancia de λ inferior a este umbral). Por consiguiente, en la máquina, cualquiera que sea, se observará una rápida convergencia aparente de esta serie hacia 100. El origen de este fenómeno es bastante sencillo: se demuestra que la serie obtenida modificando ligeramente uno cualquiera de los términos de la serie (a_n) converge hacia 100. Ahora bien, los errores de redondeo modifican inevitablemente ciertos términos, y esto a pesar de que la máquina sea muy precisa.

Las máquinas que utilizan la base 2 se hallan en los dos extremos de la gama de ordenadores

¿Cuáles son las causas de error y en qué medida pueden evitarse? Para responder a esta pregunta, es necesario comprender qué pasa exactamente en la máquina cuando hace las operaciones aritméticas y, ante todo, examinar la manera en que los números reales están representados en los ordenadores.

Supongamos, por ejemplo que se quiere representar el número π escrito con cinco cifras significativas, es decir, 3,14159. Es evidente que esto es absolutamente equivalente a escribirlo bajo la forma $3{,}14159\times10^0$, o bien $0{,}314156\times10^1$ etc. La representación en máquina llamada coma flotante se basa en esta equivalencia. Un sistema de coma flotante se caracteriza fundamentalmente por el dato de una base B y de un número M de cifras utilizadas para representar lo que se llama la mantisa. La escritura en coma flotante de un número x es de la forma:

$x = \pm x_0, x_1x_2x_3...x_{M-1}\ B^E$; el número $m = \pm x_0, x_1x_2x_3...x_{M-1}$ es la mantisa de x en base B (lo que significa que m es igual a $x_0 + x_1/B + x_2/B^2 + x_3/B^3 + ... + x_M/B^{M-1}$). El número E, exponente de B, es un entero.

En el ejemplo de π de las representaciones dadas más arriba corresponden a B=10, con M=6 y E=0 en el primer caso, y M=7 y E=1 en el segundo.

Si M vale 5 en vez de 6 o 7, π se escribe $3{,}1415\times10^0$ o $0{,}3141\times10^1$ para valores de E iguales a 0 y 1 respectivamente. Estas dos representaciones no son equivalentes, ya que la primera es más exacta que la segunda: la introducción de cero a la izquierda de las mantisas merma la precisión, ya reduce en igual medida el número de cifras significativas. Por esto, se adopta generalmente una representación llamada escritura normalizada en la que la primera cifra de la mantisa es no nulo., Evidentemente, para cero hay que adoptar una representación especial.



Además, tampoco hay que olvidar que los circuitos electrónicos utilizados para efectuar cálculos o memorizar los datos no tienen, generalmente, más que dos estados estables (correspondientes a dos valores de tensión eléctrica) que, por convención, se anotan como 0 y 1. Dado que solamente se memorizan los 0 y 1, cada una de las cifras de x_i de la mantisa de un número es escrita en base 2 incluso cuando B es diferente de 2. Por ejemplo, si B vale 10, la mantisa de cuatro cifras 3,141 estará representada por dieciséis bits (del inglés *Binary DigIT*), o cifra binaria, de la manera siguiente: 0011 0001 0100 0001, donde los dieciséis bits son las conversiones respectivas en vinario de 3,1,4 y 1. Es la representación BCD (de *binary coded decimal*).

El uso de la base 2 se ha generalizado por diversas razones. Por una parte, este tipo de representaciones aportaba una mayor simplicidad a los circuitos y a los programas de cálculo. Por otra parte, los estudios llevados a cabo sobre todo por el norteamericano Wiliam Cody[1] mientras trabajaba en el laboratorio nacional de Argonne, y por el australiano Richard Brent,[2] de la Universidad de Camberra, demostraron que era más satisfactoria desde el punto de vista de la precisión. Las máquinas que no utilizan la base 2 se hallan en los dos extremos de la gama de los ordenadores: son las calculadoras de bolsillo y algunos grandes ordenadores.

Los Redondeos: una Fuente Importante de Errores

La suma en coma flotante se ejecuta en varias fases sucesivas. Supongamos, por ejemplo, que se trabaja en base 10, con mantisas de seis cifras, y que se desea sumar $X=9{,}87654 \times 10^1$ e $Y=2{,}00071 \times 10^0$. En un primer tiempo, es necesario realizar una operación de alineado: se vuelve a escribir Y con un exponente igual a 1, o sea, $Y=2{,}00071 \times 10^1$, para obtener números con el mismo exponente. Seguidamente, se suman las mantisas: el resultado en este ejemplo es 10,076611. Finalmente, este número se vuelve a normalizar (con una sola cifra no nula antes de la coma), es decir, se escribe en la forma $1{,}0076611 \times 10^2$ y luego se redondea para guardar sólo seis cifras de mantisa, lo que da $1{,}00766 \times 10^2$ como resultado final. Por tanto, después de este cálculo, se han perdido las dos últimas cifras. Si bien la pérdida de precisión es despreciable para una sola operación, lo es mucho menos en el caso de operaciones repetitivas, cono en el caso de la suma siguiente: $1+1/2+1/3+...-1/n$, para n relativamente grande. Como la suma es conmutativa, estos términos pueden reunirse en cualquier orden. La primera idea que aparece en la mente es respetar el orden natural, de izquierda a derecha (de 1 a 1/n). Sin embargo, es preferible empezar por los términos pequeños para evitar que se conviertan en despreciables (1/n, por ejemplo, es despreciable ante $1+1/2+1/3$). De ahí la idea de calcular la suma en el orden inverso. En efecto, la tabla muestra que, así, el resultado es mucho más correcto. También es posible disponer el orden de los cálculos de manera que se guarden, para este problema, todas las cifras significativas utilizando métodos debidos a Michèle Pichat, de la Universidad de Lyon, y a William Kahan, de la Universidad de California en Berkeley.

> El fenómeno de cancelación es la principal fuente de error numérico en el cálculo científico

¿Cómo se efectúa, pues, una suma en un sistema de coma flotante? (véase el recuadro «Los redondeos: una fuente importante de errores») Evidentemente, no es posible sumar directamente las mantisas de números de exponentes distintos. De ahí la operación de alineamiento, que consiste en escribir estos números con el mismo exponente. Después de la suma de las mantisas, el resultado queda renormalizado. Finalmente, hay que redondear el resultado obtenido después de la etapa de renormalización para escribirlo en M cifras, lo que puede entrañar un ligero error.



Estos errores de redondeo se producen, por ejemplo, en la suma de dos números X e Y siendo X muy grande en valor absoluto frente a Y. Después del redondeo final, las cifras de más a la derecha de Y se pierden (véase el recuadro «Los redondeos: una fuente importante de errores»). Se trata de un fenómeno llamado de absorción. Estos errores se producen también en la multiplicación de dos números que tienen M cifras de mantisa; mientras que el resultado exacto debería escribirse con 2M cifras, el resultado que proporciona el ordenador solo tiene M: las M cifras de más a la derecha se han perdido. Aunque esta pérdida es despreciable para una sola operación, puede que no lo sea en absoluto en caso de repetición.

El otro problema relacionado con la aritmética en coma flotante se llama cancelación y se da en la resta de dos números próximos. Un ejemplo sencillo permitirá comprender de qué se trata. Supongamos que se trabaja en base 10, con mantisas de 4 cifras, y que se efectúa la resta C=A–B,

Errores Bien Acotados

En la base de la aritmética de intervalos hay la idea de acotar cada número real (por ejemplo z) con dos números máquina, uno inmediatamente inferior y otro inmediatamente superior a z (Z_1 y Z_2 respectivamente). Entonces, z es representado por el intervalo $[Z_1Z_2]$. Esto conduce a definir reglas operativas, de las cuales indicaremos aquí las mas sencillas de las cuales indicaremos aquí las más sencillas Δ designa el redondeo hacia arriba y ∇ el redondeo hacia abajo):

$$[A,B] + [C, D] = [\nabla (A+C), \Delta (B+D)]$$
$$[A,B] - [C, D] = [\nabla (A-C), \Delta (B-D)]$$
$$[A,B] \times [C, D] = [\nabla (\text{mín}\{AB, AD, BC, BD\}), \Delta (\text{máx}\{AC, AD, BC, BD\}]$$
$$[A,B]/[C,D] = [A,B] \times (1/[C,D]).$$

Los intervalos obtenidos al final de un cálculo contienen siempre los valores exactos que representan; por tanto, se está seguro de evaluar el error de cálculo. En efecto, supongamos que se desea calcular la fracción $t= 1/\sqrt{(1-x)}$, con los números escritos en base 10 con cuatro cifras de mantisa, y para x=48/49 =0,979591836... Fácilmente se comprueba que el resultado exacto es 7. En máquina, el orden de las operaciones es el siguiente: hay que calcular y=1−x, luego z=xy, finalmente t=1/z. El intervalo de partida es $X= [\nabla x, \Delta x] = [9{,}795 \times 10^{-1}, 9{,}796 \times 10^{-1}]$. Una vez realizado todo el cálculo, los números máquina inferior y superior son $6{,}983 \times 10^0$ y $7{,}003 \times 10^0$. El resultado está comprendido correctamente entre estos dos valores, pero el acotamiento continúa siendo relativamente grosero.

donde A y B son dos números en coma flotante: $A=4{,}205 \times 10^0$ y $B=4{,}199 \times 10^0$. El resultado obtenido es $C= 6{,}000 \times 10^{-3}$. Excepto si A y B son valores exactos (es decir, datos iniciales o resultados de cálculo efectuado sin error), los tres últimos ceros de la mantisa de este resultado son artificiales, tampoco es cierto que el 6 sea una cifra significativa. En efecto ¿quién sabe si el número máquina A no representa el real 4,205486 y el B el real 4,198787? Entonces, C debería valer $6{,}699 \times 10^{-3}$. De igual modo, si A representa 4,204786 y B 4,199432, el resultado real es $5{,}354 \times 10^{-3}$.

El fenómeno de cancelación es la principal fuente de error numérico en el cálculo científico. En realidad, actúa como un amplificador del error numérico preexistente: por sí mismo, no crea error. En la mayoría de los cálculos inestables, se crean errores numéricos por acumulación de errores de redondeo que luego se amplían por cancelación. Típicamente, si se calcula $(2^{60}+1)-2^{60}$ respetando el orden indicado por los paréntesis, la mayoría de los ordenadores o calculadoras de bolsillo darán 0 como resultado, en vez de 1: la suma $2^{60} + 1$ produce un error de redondeo y la resta de 2^{60} en el resultado produce una cancelación.

Si no es posible evitar los errores ¿cómo pueden estimarse? Los ejemplos anteriores muestran hasta qué punto el resultado de un cálculo puede alejarse del resultado exacto a causa de errores de redondeo. El empleo de una máquina más exacta y el respeto de ciertas reglas de programación limitan eventualmente algunos problemas, pero no los eliminan totalmente(véase el recuadro «Los redondeos: una fuente importante de errores»). Para evaluar el error de cálculo, hay dos enfoques predominantes. El primero es determinista: da un resultado cierto. Es segundo es probabilista: el resultado es sólo probable. A pesar de todo, este último se utiliza porque generalmente permite llegar a estimaciones más finas que los métodos deterministas.

La aritmética de intervalos deriva del primer enfoque. Fue particularmente estudiada en los años 1980 por Ulrich Kulisch, de la Universidad de Karlsruhe, y por Willard Miranker, del centro de investigaciones de IBM de Yorktown (Estados Unidos). El principio es sencillo. Supongamos que se quieren sumar dos números máquina X eY (los números máquina se anotan en mayúsculas; los números reales, en minúsculas). El resultado de esta suma calculado con ordenador es, o bien el número máquina inmediatamente inferior o igual al valor z de la suma real —se anota como $Z_1= \nabla(z)$—, o bien el número máquina inmediatamente superior o igual a z o $Z_2= \Delta(z)$. En vez de representar z por Z_1 o Z_2 ¿por qué no representarlo por el intervalo $[Z_1, Z_2]$ al que pertenece? La aritmética de intervalo se basa precisamente en esta idea, debida a Ramon Moore:[6] se representan todas las variables por intervalos cuyos extremos son números máquina, y se efectúan las operaciones aritméticas con estos intervalos; así, pues, el intervalo resultado contiene siempre el resultado real exacto (véase el recuadro «Errores bien acotados»). El error de cálculo se encuentra aumentado con certeza.

La Universidad de Karlsruhe puso a punto unos programas informáticos que permiten hacer fácilmente aritmética de intervalos, la cual se utiliza cuando se desea acotar de manera segura una solución. En cambio, para conocer el orden de magnitud del error cometido en una operación en coma flotante, hay que evitarla: como por naturaleza magnifica los errores, conduce muchas veces a estimaciones pesimistas.

Programas informáticos que permiten hacer fácilmente aritmética de intervalos

Hay que buscar, entonces, otros métodos, llamados de perturbación, que derivan de un enfoque probabilista.[7,8,9] La idea básica es sencilla e ingeniosa: para comprobar el efecto, en el resultado final, de las pérdidas de información por redondeo ¿no es el mejor medio simularlos y visualizar su influencia? Para ello, basta perturbar los cálculos provocando voluntariamente en cada operación un error aleatorio del orden de magnitud del error aleatorio del orden de magnitud del error de redondeo. En la práctica, el mismo cálculo perturba-

do se efectúa varias veces; los resultados obtenidos son diferentes cada vez, habida cuenta del carácter aleatorio del error. Se interpretan entonces estadísticamente los diversos resultados obtenidos. Durante mucho tiempo este enfoque solamente tuvo un carácter bastante empírico: si las perturbaciones influían poco sobre el resultado final de un cálculo, se estimaba que éste era preciso. Un punto de vista razonable pero, desafortunadamente, a veces inexacto.

Fue a partir de 1974 cuando empezaron a prepararse modelos estadísticos con el fin de sacar el mejor partido posible de los cálculos perturbados. Mencionemos, por ejemplo, el método de evaluación de error de redondeo CESTAC (*Control et Estimation Stochastique des Arrondis dans des Calculs*), de Michel La Porte y Jeans Vignes, respectivamente de la Universidad Pierre et Marie Curie y del Institut Français du Pétrole (IPF). Hasta el momento, es el único método realmente operativo.

Pero más que contentarnos con las máquinas existentes y los algoritmos de operación usuales ¿por qué no idear sistemas de representación que permitan calcular con más rapidez y más exactitud?

Para responder a esta pregunta, veamos el algoritmo de la suma en nuestro sistema de numeración en base 10. Efectuemos la suma de dos números $x_n...x_1x_0$ e $y_n...y_1y_0$ donde x_0 e y_0 son las cifras de las unidades, x_1 e y_0 son las cifras de las unidades, x_1 e y_1 las cifras de las decenas, etc. Se empieza por sumar las cifras de la derecha, x_0 e y_0. Si suma es estrictamente inferior a 10, entonces pasará a ser la cifra de la derecha del resultado; si no, ésta será igual a dicha suma menos 10, y se creará un acarreo igual a 1. Lo mismo se hace con x_1 e y_1 (teniendo en cuenta el acarreo, si lo hubiere), y así sucesivamente hasta *n*.

Un circuito de ordenador puede calcular todas las cifras de la suma al mismo tiempo

La principal particularidad de este algoritmo es la de ser intrínsecamente secuencial: el acarreo en el rango *i* depende del del rango *i*–1, etc. Esto no afecta al calculador humano: si bien somos capaces de hacer varias cosas simultáneamente (un informático diría en paralelo) —andamos a la vez que hablamos—, en cambio, somos totalmente incapaces de efectuar dos cálculos diferentes exactamente en el mismo tiempo. Por tanto, también seríamos incapaces de calcular de una sola vez todas las cifras del resultado de la suma. Así, pues, que el algoritmo utilizado implica también esta simultaneidad no es muy apremiante. Pero un circuito de ordenador sí es capaz de hacer varios cálculos a un tiempo. Por consiguiente, la utilización en la máquina de un algoritmo secuencial puede constituir un serio inconveniente. ¿No sería posible obtener todas las cifras de una suma simultáneamente? En nuestro sistema usual de escritura de los números, la respuesta es negativa: a lo sumo, puede acelerarse el cálculo de los acarreos. Por este motivo, es interesante idear sistemas de numeración en los cuales la suma pudiera efectuarse de manera totalmente paralela.

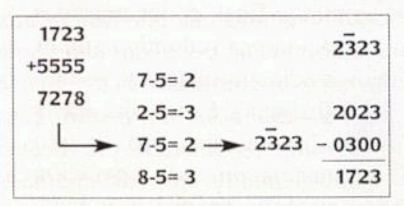
En realidad, existen varios sistemas de este tipo. Vamos a presentar uno de ellos, debido a Algirdas Avizienis, de la Universidad de California, Los Ángeles.[10] Recordemos que en el sistema usual, los números en base B (por ejemplo, en base 10), se escriben con cifras comprendidas entre 0 y B–1 (entre 0 y 9 en el caso de la base 10). Ahora bien, en 1961, Avizienis propuso escribir los números en base B con cifras tomadas entre –a y +a, donde a está comprendida (en sentido amplio) entre 1 y B–1, y donde 2a+1 es superior o igual a B. Por ejemplo, en base 10, y utilizando cifras comprendidas entre –7 y 7 (a=7), el número 1723 puede escribirse $2\bar{3}23$, en el que las cifras con una barra en la parte superior se consideran negativas ($\bar{3}$ = –3). En efecto, $2\bar{3}23$ es igual a $2 \times 1000 + \bar{3} \times 100 + 2 \times 10 + 3 \times 1$. Sin embargo, puede comprobarse fácilmente que, en este sistema 1723 puede también escribirse $173\bar{7}$, $2\bar{3}3\bar{7}$, $2\overline{277}$ (véase el recuadro «Del sistema decimal usual a los sistemas redundantes»).

Por consiguiente, algunos números tienen varias representaciones. Por esto, los sistemas de Avizienis suelen calificarse de redundantes. Se les llama también sistemas de cifras negadas. Permiten representar todos los números enteros por cuanto que 2a+1 es superior o igual a B. Por ejemplo, si a es superior o igual a 5, podrán representarse todos los números en base 10. Las conversiones entre nuestro sistema de numeración habitual y los sistemas de Avizienis se efectúan muy fácilmente.

En su artículo de 1961, Avizienis proponía un algoritmo que permite efectuar sumas de manera totalmente paralela (es decir, calculando todas las cifras de una suma simultáneamente), a condición de que 2a sea superior o igual a B+1, y de que a sea inferior o

Del Sistema Decimal Usual a los Sistemas Redundantes

Para convertir un número representado en base 10 según la escritura usual en la escritura en «cifras negadas» (a la izquierda), con cifras tomadas entre –5 y 5, en primer lugar se añade 555...5 (tantos «5» como cifras tiene el número a convertir) al número que se quiere convertir y luego se resta 5 a cada cifra de la suma (excepto el último si esta suma comporta una cifra de más que el número de partida). Para la operación inversa (a la derecha), se descompone el número a convertir en dos cantidades que no tengan cifras positivas: la primera está constituida únicamente por cifras positivas del número de partida —insertando ceros en los lugares de las cifras negativas— y la segunda está constituida por los opuestos de las cifras negativas. Sólo hay que restar estas dos cantidades para obtener la escritura usual del número de partida.



Para más Información Sobre el Algoritmo de Avizienis



El profesor Algirdas Avizienis es, entre otras cosas, profesor emérito de la Universidad de California, Los Ángeles (Estados Unidos). (Foto J. Jamulaitis).

El algoritmo de Avizienis permite efectuar sumas de manera totalmente paralela. Se basa en una representación particular de los números: cada número en base B se escribe con cifras comprendidas (en el sentido amplio) entre a y –a, donde 2a es superior o igual a B+1 e inferior o igual a B–1 (a condición de que B sea diferente de 2; por tanto, el algoritmo no es compatible con la base 2).

El resultado $s_{n+1}...s_1\ s_0$ de la suma de dos números $x_n...x_1x_0$ e $y_n...y_1y_0$ en base B se obtiene calculando en paralelo la suma w_i+t_i, para i de 0 a n+1, donde w_i y t_i son calculados en paralelo de la manera siguiente:

$$t_{i+1}=\begin{cases} -1\ s_i\ x_i+y_i \le -a \\ +\ s_i\ x_i+y_i \le +a \\ 0\ si\ -a+1 \ge x_i \le a-1 \end{cases}$$

$w_i=x_i+y_i-Bt_{i+1}$ y por convención $w_{n+1}=t_0=0$

Los términos t_{i-1} desempeñan una función similar a la de los acarreos en una suma normal, pero con la diferencia de que no hay ninguna dependencia entre t_i y t_{i+1}, lo que permite el paralelismo. Es importante recordar que, utilizado este algoritmo, un circuito de ordenador puede calcular todas las cifras de la suma al mismo tiempo. Por ejemplo, supongamos que se desea sumar los dos números escritos en base 10 con a=6: $1\bar{2}\bar{4}5$ (=765) y $\bar{2}3\bar{3}2$ (= –2328) (donde $\bar{2}$ = –2).

El resultado obtenido, $\bar{1}\bar{6}4\bar{3}$ (= –1563) es igual a la suma de 765 y –2328.

En base 2, y utilizando las cifras –1, 0 y 1, también es posible concebir algoritmos completamente paralelos de suma que no son muy diferentes del algoritmo descrito. Estos algoritmos permiten sumar dos números escritos en *n* cifras (para *n* cualquiera) en un poco más de tiempo que el necesario para que un sumador clásico sume dos números de 2 cifras.

i	3	2	1	0
x_i	1	$\bar{2}$	$\bar{4}$	5
y_i	$\bar{2}$	$\bar{3}$	$\bar{3}$	2
x_y+t_{i+1}	-1	-5	-7	7
$t_i + 1$	0	0	-1	1
w_i	-1	-5	$\bar{3}$	-3
s_i	$\bar{1}$	6	4	3

igual a B–1. Estas condiciones son algo más restrictivas que las descritas anteriormente. Este algoritmo aprovecha la redundancia del sistema de numeración utilizado para evitar la propagación de arrastres a medida que se hacen sumas parciales. De este modo, un circuito de ordenador puede calcular todas las cifras de la suma al mismo tiempo (véase el recuadro «Para más información sobre el algoritmo de Avizienis»).

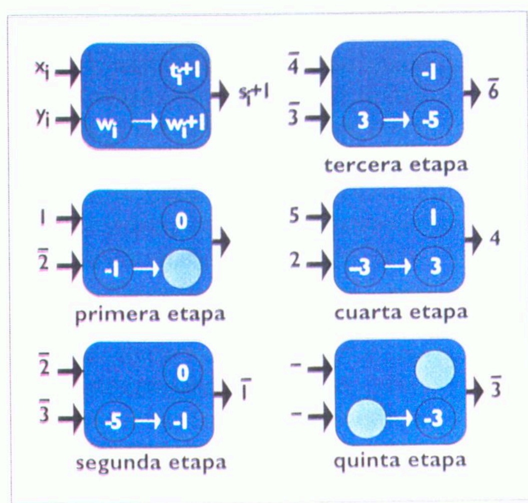
La redundancia de los sistemas de numeración de Avizienis no está exenta de inconvenientes, pero es necesaria: en su tesis, Christophe Mazenc, de la Escuela Normal Superior de Lyon, demostró que, bajo ciertas condiciones, un sistema de numeración que permita efectuar sumas de manera totalmente paralela es necesariamente redundante. [11]

Si con estos sistemas es posible efectuar sumas mucho más rápidamente ¿por qué no se conocen y utilizan más? Pues porque presentan algunos inconvenientes. En primer lugar, necesitan mucha más memoria y conexiones en los circuitos. Si por ejemplo, se quieren representar todos los números enteros comprendidos entre –2048 y +2047, bastarán 12 bits en la notación usual (no redundante) en base 2, mientras que serán necesarios 24 si se emplea la base 2 y las cifras –1,0 y 1 (2 bits para representar cada cifra), y 18 si se utiliza la base 4 y cifras comprendidas entre –3 y 3. Por otra parte, aunque la suma es más rápida, las comparaciones entre dos números resultan más lentas: es más difícil comparar dos números que admiten varias representaciones. Además, la conversión de un número de un sistema redundante a un sistema no redundante (de igual base) requiere tiempo, el de una suma en sistema usual (véase el recuadro «Del sistema decimal usual a los sistemas redundantes»).

Los circuitos que utilizan la aritmética llamada en línea constituyen un ejemplo de arquitectura especializada

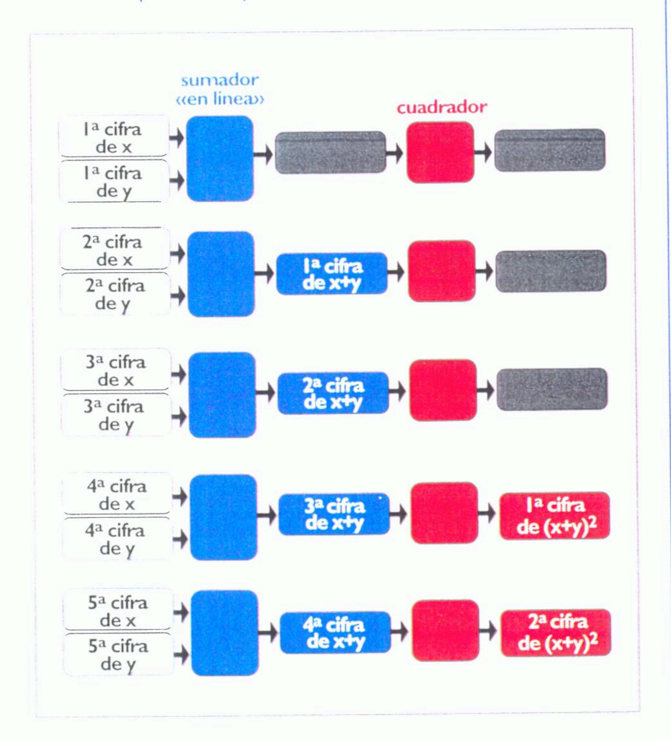
En consecuencia, y aunque teóricamente sea posible, la construcción de un ordenador (es decir, una máquina universal capaz de trata problemas muy diferentes) que se base en la utilización de un sistema redundante de escritura de los números no es realista, al menos en el estado actual de los conocimientos y de las tecnologías. [12] En cambio, se perfilan dos campos peculiares de aplicación. Uno de ellos corresponde al de las arquitecturas especializadas, es decir, de máquinas dedicadas a una aplicación particular (control de procesos, tratamiento numérico de la señal, etc.), que no tienen que efectuar las operaciones que los sistemas redundantes hacen mal. El segundo es el de los operadores de tipo caja negra. Los llamo así porque únicamente su diseñador sabe que ponen en práctica un sistema de numeración especial. Estos operadores, que están integrados en un ordenador, reciben los datos y facilitan su resultado en un

Del cálculo «en línea» al «pipe line»



Los sistemas redundantes de representación de los números permiten aplicar un modo de cálculo particular llamado en línea, que se esquematiza aquí. Reexaminemos el algoritmo de suma de Avizienis, presentado en el recuadro «Del sistema decimal usual a los sistemas redundantes», y supongamos que recibimos las cifras de los datos una por una, de izquierda a derecha (al principio se reciben x_n e y_n, después x_{n+1} e y_{n+1}, después x_{n-1} e y_{n-1}, etc.). A partir de x_n e y_n se puede deducir t_{n+1} y w_n. En la etapa siguiente, x_{n-1} e y_{n-1} permiten calcular t_n y w_{n-1}. Entonces se puede suministrar la primera cifra del resultado partiendo de la izquierda, s_n que es igual a $w_n + t_n$. En la etapa siguiente, el proceso es el mismo con x_{n-2} e y_{n-2} y así sucesivamente: de los valores x_i e y_i se deducen los de t_{i+1} y w_i y, por consiguiente, s_{i+1} (según el algoritmo de Avizienis). En el esquema de abajo se han representado las diferentes etapas de la suma, a partir del ejemplo expuesto (gráfico) en el recuadro «Del sistema decimal usual a los sistemas redundantes», es decir, la suma $1\bar{2}\bar{4}5$ (=765) y $\bar{2}\bar{3}\bar{3}2$(=−2328). El interés de este modo de cálculo, llamado en línea, es que las primeras cifras del resultado quedan disponibles para el cálculo siguiente antes de que se haya terminado el cálculo: s_n queda disponible para la operación siguiente en el momento en que se reciben x_{n-1} e y_{n-1}, etc. Si se hiciesen circular las cifras de derecha a izquierda, se podrían construir operadores aritméticos que tendrían un comportamiento similar para la suma y la multiplicación (y, por tanto, sin tener necesidad de un sistema redundante de representación de los números), pero la mayoría de los demás cálculos serían imposibles. En aritmética en línea, se saben construir estos operadores para casi todas las funciones matemáticas usuales (división, raíz cuadrada, seno, coseno, logaritmo, exponencial, etc.). Este algoritmo de suma permite obtener la (i-1)^{ava} cifra (partiendo de la izquierda) de la suma de los dos números desde que se han entrado en el operador las i^{avas} cifras de estos números. Se dice que la suma es de retardo 1: un operador de retardo δ si proporciona la cifra (i-δ)^{ava} del resultado después de la recepción de las i^{avas} cifras en sus entradas. Por ejemplo, se saben hacer multiplicaciones «en línea» con un retardo igual a 2. La figura adjunta presenta lo que sucede al calcular $(x+y)^2$ en aritmética «en línea» con un sumador de retardo 1 y un cuadrado (es decir, un operador que eleva al cuadrado su entrada) de retardo 2. Al recibir la i^{ava} cifra de x e y, se obtiene la (i-1)^{ava} cifra de x+y de la que se deduce la (i−3)^{ava} cifra de $(x+y)^2$.

Por tanto, la suma y la elevación al cuadrado se realizan al mismo tiempo y se pueden utilizar las primeras cifras de x+y antes de que se haya terminado el cálculo de x+y. Esta



técnica, bien conocida por los informáticos, se llama «pipe line» o «tubería». Utilizando esta técnica, el retardo global de una secuencia de cálculo será igual a la suma de los retardos de los operadores utilizados.

sistema de numeración clásica, pero su cálculo interno lo efectúan en un sistema redundante de escritura de los números. Los circuitos que utilizan la aritmética llamada en línea constituyen un ejemplo de arquitectura especializada. El pionero de este modo de cálculo es Milos Ercegovac, profesor de la universidad de California, Los Ángeles.[13] Los sistemas de cálculo en línea reciben sus datos y facilitan los resultados cifra tras cifra, empezando por las más significativas. Ya no hay simultaneidad en el cálculo de las cifras del resultado, pero esto queda compensado por una propiedad interesante: en el encadenamiento de varias operaciones, un operador aritmético puede trabajar con las primeras cifras de un número mucho antes de que el operador precedente haya terminado su cálculo. Los informáticos llaman a este proceso —muy parecido al principio de trabajo en cadena— una *pipe-line* (o «tubería») a nivel de la cifra (véase el recuadro «Del cálculo «en línea» al «*pipe-line*»). De este modo, es posible efectuar muy rápidamente secuencias complejas de cálculos. Los primeros algoritmos en línea para la suma, la multiplicación y la división datan de 1977. En cuanto a los de otras funciones, algunos se construyeron en los años 1980, mientras que los hay que son todavía objeto de un campo de

investigación activo. En el laboratorio de informática del paralelismo, de la Escuela Normal Superior de Lyon, estamos trabajando actualmente en el diseño de algoritmos y circuitos para el cálculo en línea de las funciones seno, coseno, exponencial y logaritmo. La aritmética en línea puede conducirnos a dominar mejor la precisión: en una cadena de cálculos, el usuario recibe las cifras del resultado una a una y puede prolongar el cálculo hasta que haya obtenido toda la precisión que desea. En Francia, estas vías han sido exploradas por Christophe Mazenc (citado más arriba) y Valérie Ménissier, del INRIA (*Institut National de Recherche en Information et en Automatique*), de Roquencourt.[11,14] Por ejemplo, Valérie Ménissier ha escrito una biblioteca de programas de manipulación de números reales, de precisión arbitrariamente grande, que se ha integrado en un lenguaje de programación (el lenguaje CAML).

En cuanto a los operadores de tipo caja negra, sólo son eficaces si el volumen de cálculo a efectuar es importante respecto a una suma con propagación de arrastre (suma según el algoritmo usual), ya que la conversión de los datos, en la entrada y en la salida del operador, necesita el tiempo de dicha suma. Hoy, la mayor parte de los multiplicadores (es decir, los circuitos que efectúan multiplicaciones en los ordenadores) utilizan para sus cálculos internos una representación llamada *carry-save*, que es un sistema redundante de escritura de los números en base 2, con las cifras 0,1 y 2. Pero en 1985, Naofumi Takagi, Hiroto Yasuura y Shuzo Yajima, de la Universidad de Kioto, propusieron un nuevo multiplicador de gran eficiencia, en el cual los cálculos se hacen en base 2 con las cifras –1, 0 y 1.[15] Es un multiplicador de este tipo el que se implantó en uno de los coprocesadores matemáticos construido por la empresa Cyrix, que puede colocarse en un microordenador para acelerar los cálculos. Posteriormente, han aparecido numerosos operadores de caja negra para la división.

Por nuestra parte, nos interesamos también muy de cerca en el diseño de uno de estos operadores para calcular las funciones matemáticas de base (seno, coseno, exponencial, logaritmo, etc.).[16] Acabamos de poner a punto un algoritmo adaptado a este tipo de operadores. Presenta la particularidad de utilizar solamente sumas y divisiones por potencias de 2.

Una potencia de 2 es un número de la forma $2\times2\times2\times...\times2$: dividir por un número de éstos cuando se trabaja en base 2 es tan sencillo como dividir por 10, 100, o 1000 cuando se trabaja en base 10: consiste en un simple desplazamiento de la coma. Nuestro algoritmo permite calcular rápidamente, entre otras, las funciones seno, coseno, logaritmo, exponencial, arco tangente y raíz cuadrada. Debido a sus cualidades(rapidez, gran número de funciones calculables), podría emplearse en las pequeñas calculadoras y en los procesadores en coma flotante.

La aritmética en coma flotante se utiliza para proyectar presas, puentes, automóviles, etcétera

Este artículo tenía por objeto presentar dos campos importantes de la aritmética de los ordenadores: la evaluación de errores y el cálculo rápido gracias al empleo de la aritmética redundante. Si de todo ello el lector deduce que no hay que depositar jamás una confianza ciega en los resultados facilitados por un ordenador y que la implantación rápida de una operación tan elemental como la suma compete todavía al campo de la investigación, me daré por absolutamente satisfecho. Los ejemplos numéricos dados al principio del artículo son voluntariamente lúdicos: son simples ejemplos de escuela, pero no por ello hay que olvidar que ponen de manifiesto un problema real. Una larga serie de cálculos puede conducir a un resultado numérico muy impreciso y las consecuencias de tal imprecisión es posible que sean graves: los cálculos en coma flotante se hacen en misiles aviones, etc. La aritmética en coma flotante se utiliza para proyectar presas, puentes, automóviles, etc. En su tesis doctoral, Douglas Priest, de la universidad de California, Berkelley, cita el caso de un misil *Patriot* que, durante la guerra del Golfo, no consiguió interceptar un misil iraquí, debido a un error numérico. El gran matemático Arnold Householder decía que tenía miedo a subir a una avión desde que sabía que estos aparatos se proyectan utilizando la aritmética en coma flotante. Naturalmente, se trataba de una bromea, pero esperemos que el futuro no habrá de darle la razón. ■

JEAN-MICHEL MULLER está a cargo de las investigaciones CNRS en el Laboratorio de informática del paralelismo en le Escuela Normal Superior de Lyon.

(1) W.J. Cody, *IEEE Trans. On Computers, vol. C-22* nº 6, 598, 1973
(2) R.P. Brent, *IEEE Trans. On Computers, vol. C-22*, nº6, 601, 1973.
(3) M. Pichat, *Num. Math., 19*, 400, 1972.
(4) W. Kahan, *Paradoxes in Concepts of Accuracy*, actas del *Joint Seminar on Issues and Directions in Scientific Computation*, U.C. Berkeley, 1989.
(5) R.E. Moore, *Interval analysis*, Prentice Hall, Englewoos Cliffs, Nueva York, 1963
(6) U.W. Kulisch y W.L. Miranker, *SIAM Rev., 28*, 1, 1986
(7) M. La Porte y J. Vignes, *Number. Math., 24*, 39, 1975
(8) J. Vignes *AFCET Interfaces, 54*, 3, 1987.
(9) F. Chatelin, *Analyse statistique de la qualité numérique et arithmétique de la résolution approchée d'équation par calcul sur ordinateur*, Estudio F-133, Centro científico IBM Francia, París, abril 1988.
(10) A. Avizienis, *IRE Transacions on Electronic Computers, 10*, 398, 1961.
(11) C. Mazenc, *Systèmes de représentation des nombres et arithmétique sur machines parallèles*, Tesis doctoral, Universidad Claude Bernard de Lyon y Escuela Normal Superior de Lyon, diciembre 1993.
(12) J. Duprat y J.M. Muller, *Technique et Science informatique, vol. 10, nº3*, 1991.
(13) M.D. Ercegovac y K. Trivedi, *IEEE Transactions on Computers, vol C-26 nº7*, 1977
(14) V. Ménissier, *Arithmétique exacte*, Tesis doctoral, Universidad de París VII, diciembre 1994.
(15) N. Takagi *et al., IEEE Transactions on Computers, vol C-34, nº9*, setiembre 1985.
(16) J.C. Bajard *et al, IEEE Transactions on Computers, vol C-43 nº8*, agosto 1994.

Para Más Información:

D. Goldberg, *What every computer scientist should know about floatintg-point arithmetic*, ACM Computing Surveys, Vol. 23 nº1, marzo 1991
J.M. Muller, *Arithmétique des ordinateurs*, Coll. Études et Recherches in Informatique, Massin, Francia, 1989.
J.L. Hennessy y D.A. Patterson, *Architecture des Ordinateurs*, edición francesa por Daniel Etiemble y Michel Israel, Ediscience International, 1994
A.R. Ormondi, *Computer Arithmetic Systems*, Prentice Hall International series in Computer Science, Prentice Hall, 1994
J. Wallis, *Improving Floating-Point Programming*, J. Wiley & Sons, 1990.

Los números, esencia de las cosas

Bernard d'Espagnat

Al matematizar la realidad, la física contemporánea parece invitarnos a acercarnos al punto de vista de Pitágoras, para quien los números son la esencia de las cosas. Pero hay que renunciar al ideal einsteiniano de una realidad en sí misma totalmente cognoscible gracias a las matemáticas.

Como bien dicen los informáticos y los vendedores de material de alta fidelidad, vivimos en una época totalmente numérica. El número se ha convertido en un material prodigiosamente eficaz. Si es hasta tal punto eficaz, ¿no será porque se encuentra en el corazón mismo de lo existente, más incluso que el «lugar» y la «cosa»?

Realismo próximo. La cuestión vale la pena de plantearse y no es ilegítimo hacerlo así. A la escala más fina, las leyes que rigen la materia son las de la física cuántica y el adjetivo mismo «cuántico» designa una discontinuidad a la que la idea de lo numérico remite inmediatamente. Pero vayamos por partes. Hoy por hoy, la técnica de los componentes electrónicos carece del grado de finura más allá del cual predominan los comportamientos típicamente cuánticos. Por otra parte, pese al vocabulario, la discontinuidad no es el aspecto fundamental de las leyes cuánticas. Es más bien una de sus consecuencias. Como tampoco los números constituyen, ni mucho menos, la totalidad de las matemáticas.

La invención del ordenador es comparable a la de la rueda o, mejor dicho, a la del cubo de la rueda, que es su «principio activo». Pero en la naturaleza no hay cubo. ¿Hay realmente matemáticas? ¿Son ellas la esencia última de la naturaleza? O por el contrario, ¿están ahí porque las hemos puesto nosotros? La eficacia de lo numérico ni lo prueba ni lo desmiente. Para averiguar cuál es la situación, lo adecuado es preguntar a la propia física y no a las técnicas derivadas de ella.

La física moderna nació con Galileo y en ella las matemáticas desempeñaron de inmediato un papel fundamental. ¿Quién no conoce la afirmación de Galileo según la cual el lenguaje de la naturaleza está escrito en lenguaje matemático y quien pretenda leerlo debe primero aprender dicho lenguaje? Pero esta tesis, opuesta a la ciencia cualitativa de Aristóteles, suscita a su vez varias cuestiones, porque hay distintas maneras de entenderla. La que mejor concuerda con el sentido común y con la física clásica (la física desarrollada hasta comienzos del siglo XX) consiste en distinguir entre la «naturaleza» de las cosas y las «relaciones» entre las cosas y en considerar que las matemáticas, que son fundamentalmente ciencia de las relaciones, intervienen no en la definición de la propia naturaleza sino en la de sus relaciones mutuas.

> No es por medio de imágenes de la vida corriente como se puede explicar, como hace la relatividad, que el espacio se transforma en tiempo

Concebida así, la física concuerda bastante bien con las concepciones del llamado «realismo próximo». Éstas (tenidas por indudables en el siglo XVII e incluso más tarde) consisten en considerar que la mente humana dispone innatamente de «nociones claras y distintas» (según la expresión de Descartes), es decir, de conceptos —el de espacio (euclídeo), el de forma, el de objeto localizado, etc.—, cuya «corrección» garantiza la correspondencia con la realidad. El objetivo del físico adepto a estas concepciones debe consistir en explicar los fenómenos que observa por medio de dichos conceptos y sólo de ellos. Una versión muy suavizada del realismo próximo consistiría en considerar «evidentes» un gran número de conceptos (entre los cuales los de vida, de mente, etc.). Pero la versión que históricamente se impuso, el «cartesianismo» o «mecanicismo», es, por el contrario, una versión «dura», porque admite sólo un pequeño número de conceptos (para Descartes sólo eran admisibles las figuras, las magnitudes y los movimientos). Ello obliga, en contrapartida, a hacer gran uso de las matemáticas para describir las relaciones que permiten que los objetos correspondientes a dichos conceptos se agreguen en estructuras complejas, interactúen y así sucesivamente. Como es sabido, este segundo rumbo es el que tomó principalmente la física clásica. Fue el auténtico «camino real».

Explicaciones intuitivas. Esta vía no excluye en absoluto la «construcción de conceptos». Al contrario, como subrayaba Bachelard, la mayoría de los conceptos utilizados en física clásica son construcciones de la mente humana. Las nociones de punto material, de interferencias entre dos ondas, etc., son construcciones. Pero hay un criterio cuya aplicación nos demuestra que su in-

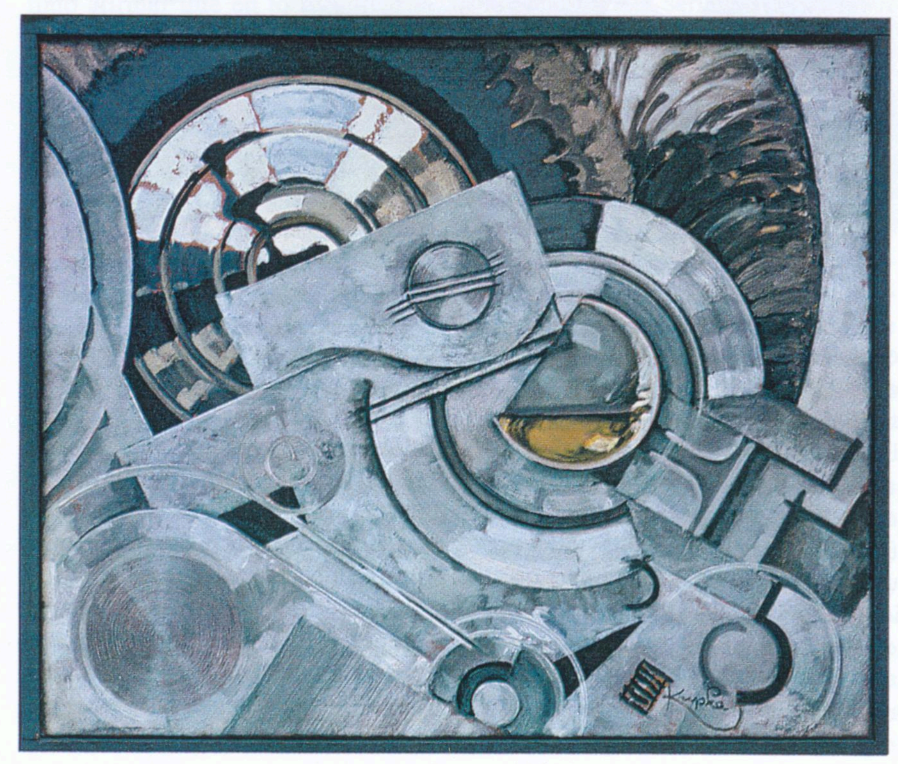
troducción no nos aleja del realismo próximo. Se trata de que cabe explicar intuitivamente el significado de dichos conceptos por medio de imágenes tomadas de la vida real y de pasos al límite (el punto material que evoca granos de arena cada vez más finos, las interferencias que esbozan la imagen de un dique portuario en el que se han abierto dos orificios). Ello pone de manifiesto que en el realismo próximo el papel de los números y las matemáticas, aunque esencial en la práctica, es secundario desde el punto de vista conceptual. Su papel es de mero enlace y el papel principal es atribuido a las nociones claras y distintas, es decir, a los conceptos simples y «evidentemente adecuados».

Los límites del sentido común. Ahora bien, el punto esencial de este artículo es que las imágenes de la vida cotidiana (aunque estén muy idealizadas) son incapaces de explicar, como hace la teoría de la relatividad, que el espacio se transforme (parcialmete) en tiempo (y viceversa) cuando se cambia de sistema de referencia o que el movimiento puro se transforme en cosas (pensemos en los fenómenos de creación de partículas en los choques de altas energías).

En este sentido, el nacimiento de la relatividad constituye una especie de ruptura. Los conceptos familiares, las nociones de «sentido común» tan caras a Descartes, ya no bastan y a veces llegan a ser inadecuadas. Dicho de otro modo, el realismo próximo resulta falso. ¿Por qué ha sido posible esta ruptura, y, sobre todo, por qué ha sido fecunda? Es lo que vamos a examinar ahora.

> La mayoría de los científicos parecen dar por sentado que la investigación trata de desvelar la realidad y lo logra bastante bien

La respuesta hay que buscarla en otra interpretación de la célebre aserción de Galileo, una interpretación mucho más radical que la primera y que recuerda la antigua exclamación atribuida a Pitágoras: «¡Los números son la esencia de las cosas!». Según esta interpretación, el libro de la naturaleza está necesariamente escrito en lenguaje matemático porque las cosas no son «cosas» propiamente dichas (objetos masivos y localizados) sino representaciones de grupos de transformaciones* o soluciones de ecuaciones. Las matemáticas, pues, no pueden limitarse a expresar relaciones. Lo que traducen es la esencia misma de la realidad.



«En la naturaleza no hay nada semejante al cubo de una rueda. ¿Hay allí matemáticas?» Frank Kupka, 1927, L'Acier boit nº 2.
(Foto © Giraudon by ADAGP 1999 Marseille, Musée Cantini)

Esta matematización (o «descosificación») de la realidad es fundamental en la física contemporánea. En sus memorias, Heisenberg relata que cuando era estudiante consideraba inverosímil y ridícula la explicación que su manual de química daba del fenómeno de la valencia (donde se hablaba de «átomos ganchudos»[1]). También recusaba por arbitrarias todas las imágenes de los átomos propuestas por los libros de divulgación. Pero al no encontrar alternativas, el joven alemán se encontraba conceptualmente estancado. Heisenberg explica que fue la lectura del *Timeo* de Platón lo que le hizo entrever la salida, sugerida por el hecho de que, renunciando a describir como cosas las «partes más pequeñas de la materia», el autor del diálogo remitía para ello, aunque en términos ambiguos, a las matemáticas. Y efectivamente, Heisenberg, unos años después, fundó sobre unas bases matemáticas puramente abstractas la teoría atómica actual, sacando así a la física de lo más pequeño del callejón sin salida en que se hallaba. Con mayor generalidad, los datos fácticos de la física del siglo XX han impuesto una rotunda superación del realismo próximo y han conferido un papel eminente a las matemáticas.

Este cambio es radical. Antaño, por ejemplo, Descartes, refiriéndose a la estructura de la materia, hablaba de los «*tubos o resortes que causan los efectos de los cuerpos naturales*», y precisaba que «*eran demasiado pequeños para ser percibidos por nuestros sentidos*». En el siglo XVIII, todavía nadie habría considerado superable la regla cartesiana del uso exclusivo de nociones «claras y distintas» proporcionadas por el entendimiento y el sentido común, que por lo tanto eran «verdaderas». El que semejante superación, realizada siguiendo métodos adecuados, no desemboque en vagos esoterismos sino en teorías experimentalmente confirmadas es un descubrimiento capital. Advirtamos de paso que el descubrimiento lo debemos casi exclusivamente a la física. Las demás ciencias, biología, etc., recurren sin duda a nociones extremadamente complejas; no obstante, en el uso que hacen de ellas, casi todas pueden vincularse a unos conceptos constitutivos mucho más simples, de la familia del realismo próximo.

El poder de la mente humana. Así, la física permite superar el marco de las nociones primeras de «sentido común» y de las construcciones vinculadas a estas últimas (el marco del realismo próximo). La posibilidad de semejante superación constituye un testimonio impresionante del poder de la mente humana. Pero no es motivo suficiente para envanecerse. Una excesiva exaltación de las capacidades humanas podría llevarnos a una cruel desilusión.

¿Levantar el velo? La cuestión delicada no es tanto el realismo próximo como el realismo en general. Descartes (¡otra vez él!) fue el primero en suscitarla. El filósofo francés se preguntaba por la posibilidad de que,a fin de cuentas, nuestros sentidos nos engañaran (incluso ayudados por nuestros instrumentos), lo cual le llevaba a inquirir acerca de la posibilidad de que nuestra ciencia pudiera esperar alcanzar, más allá de lo simplemente observable, el conocimiento de lo real. Es la famosa «duda cartesiana», bien conocida por los filósofos. Pero éstos saben perfectamente que Descartes dio en definitiva una respuesta positiva a esta cuestión, superando la famosa duda. Los avances de la física confirmaron a sus sucesores en este punto de vista. En realidad, la inmensa mayoría de los científicos, sin haber pensado siquiera en el asunto, parecen considerar evidente que el estudio pretende desvelar la realidad y lo logra bastante bien. Cierto es que el objetivo no puede alcanzarse por el camino del realismo próximo tan caro a Descartes. Pero *a priori* nada impide conjeturar que la meta podría alcanzarse por medio de la idea de las «matemáticas, esencia de las cosas» considerada antes. No es impensable que la física matemática haya descubierto finalmente los conceptos correctos —espacio curvo en vez de espacio plano, $E = mc^2$ en vez de energía y masa—, los que reflejan fielmente la estructura de lo real en sí. Tal era la esperanza de Einstein, como la de otros muchos físicos de nuestro tiempo.

Para levantar el velo, no obstante, sería preciso que todas las leyes naturales, tanto las de la física macroscópica como las de la física atómica y subatómica, pudieran formularse así. Pero en este punto la física cuántica nos envía una señal de alarma que no se puede pasar por alto. Algunos de los axiomas de esta disciplina, tal como se la enseña, se presentan bajo la forma de reglas de predicción del tipo «si se hace esto se observará aquello». En otras palabras, las matemáticas que esta disciplina pone en juego son unas «matemáticas no ontológicas»: los algoritmos que introducen reflejan con precisión nuestra aprehensión humana de lo real y nuestra capacidad de acción pero no pueden considerarse como traducciones de los elementos estructurales de la realidad «en sí misma».

¿Es definitivo este estado de cosas? La pregunta constituye uno de los aspectos de una problemática hoy en día muy abierta que consiste en preguntar si las interpretaciones tradicionales de la mecánica cuántica pueden y deben ser sustituidas. Desde hace tiempo se llevan a cabo importantes esfuerzos por conseguir una teoría «ontológicamente interpretable», según reza la fórmula consagrada. Y la búsqueda prosigue. Casi cada año surge una propuesta nueva en este sentido. Los autores de algunas de ellas llegan incluso a escribir obras para el gran público donde —no sin cierta petulancia— presentan su enfoque como la «interpretación moderna» de la mecánica cuántica.

Pero la realidad no es tan optimista. El obstáculo principal se llama «no separabilidad». Se trata de que toda teoría que pretenda describir la realidad tal cual es hasta la escala microscópica y que no restrinja artificialmente el alcance admitido del término «verdad» ni viole ciertos datos experimentales debe admitir ciertas influencias instantáneas a distancia que no decrecen con la distancia u otros efectos equivalentes (violación de la causalidad local).[2] Es el teorema de Bell, que se aplica no a ciertas teorías sino a todas ellas, incluida la mecánica cuántica, tan pronto como se les imponen las condiciones que acaban de exponerse. El teorema de Bell nos obliga a formular una concepción de lo real —si es que realmente hay una realidad cognoscible— totalmente distinta de la que habíamos imaginado. Por ello, la duda cartesiana vuelve a hacer acto de presencia[1]

Afortunadamente, existe otro teorema, enunciado por primera vez por mí mismo[3] y demostrado luego por otros, que demuestra que tales influencias no pueden propagar señales, por lo que no pueden utilizarse de ninguna manera... Un resultado muy reconfortante en la práctica. Pues podemos hacer «como si» estas influencias no existieran y seguir utilizando las reglas universales de predicción de la mecánica cuántica sin preocuparnos por el problema así suscitado. Si nuestra postura filosófica es el «idealismo integral», podemos incluso negar toda realidad independiente a lo humano y por tanto también las influencias en cuestión.

De ahí a sentar pura y simplemente, sin

fidelidades idealistas ni reticencias ante la idea de realidad, que no hay ninguna influencia instantánea a distancia o efecto semejante hay, por supuesto, un abismo. Curiosamente, es un abismo que franquean alegremente algunos científicos en libros para el gran público. Por ejemplo, en *El quark y el jaguar*, Murray Gell-Mann, sin expresar aquellas fidelidades o reticencias, sin siquiera poner mínimamente en guardia, afirma pura y simplemente la inexistencia de toda influencia instantánea o efecto similar.[3] Hay que tratar, por supuesto, de explicar la presencia en dicho libro de tal alegato, sorprendente para el especialista. Por ejemplo, se podría intentar justificarlo sobre la base de la imposibilidad, mencionada antes, de enviar señales. El alegato podría defenderse sólo si hubiera que tener en cuenta los aspectos pragmáticos de las cosas: las aplicaciones.

El teorema de Bell nos obliga a hacernos una idea de la realidad totalmente distinta de la que habíamos imaginado

Pero en el campo del saber no todo se reduce a la acción. Por tanto, es legítimo afirmar que en lo tocante a la verdadera naturaleza del conocimiento que el hombre adquiere de la realidad los autores de aserciones como la anterior dan una información falaz a sus lectores. Y ello tanto más cuanto que estos últimos, siempre que no hayan sido orientados hacia una filosofía diferente, se adhieren a la idea universalmente difundida de que la ciencia, si tiene éxito, describe la realidad tal como es. Estamos ante un tipo de problema muy grave que debe afrontar la divulgación de la física y del que autores y lectores deben tomar conciencia si quieren evitar la desinformación. Qué duda cabe de que, para llevar a cabo su cometido, el autor que escribe para el no especialista debe expresarse en un lenguaje accesible: debe tomar como punto de partida nociones que son obvias para el lector. Pero por desgracia estas nociones se refieren a las concepciones del realismo próximo, el cual, como hemos visto, está superado en muchos aspectos.

Una divulgación engañosa. El «caso de conciencia» es, pues, cruel. Hasta una época muy reciente (y todavía hoy en algunos países), los autores «serios» daban prioridad absoluta al deseo de no desorientar al lector presentándole ideas demasiado extrañas para él. Se le hablaba, por tanto, de las «partículas» de la teoría cuántica como si se tratara de corpúsculos localizados en todo instante, y otras cosas del mismo estilo. Se pretendía así poner coto al esoterismo (bastante temible, es verdad), que engendra con frecuencia una comprensión imperfecta de las novedades. Pero el reverso de la medalla consistía en que al proceder así se ocultaba el mensaje esencial de la física, a saber, que esta disciplina, aunque siendo estrictamente racional, es radicalmente incompatible con el realismo próximo o cualquier idea semejante.

Cubos matemáticos. Las prácticas de esta clase siguen siendo moneda corriente. Así, los físicos que reconocen en privado que las teorías que proponen no desvelan una realidad totalmente independiente de lo humano evitan casi siempre que esta verdad aflore en sus escritos. Por ejemplo, si tuviera algún reparo que formular al excelente artículo, rico en contenido informativo, publicado en 1995 por Roland Omnès, sería precisamente la existencia de una tal laguna.[II]

Por otra parte, ninguna de las teorías auténticamente realistas que reproducen las predicciones de la mecánica cuántica (y que por tanto son no esperables, hay varias en el mercado) ha logrado superar todas las dificultades (sobre todo la conciliación con la teoría de la relatividad). Ninguna, por tanto, ha elaborado una concepción que pueda ser universalmente aceptada. Y hay buenas razones cuantitativas para ser escéptico en cuanto a la posibilidad de lograrlo.

En tales condiciones, parece razonable renunciar al ideal einsteiniano de una realidad en sí totalmente cognoscible por medio de conceptos matemáticos y cuyos elementos («elementos de realidad»), gracias a estos conceptos, queden descritos «tal como son». En cambio, parece no sólo posible sino también bastante verosímil que que las matemáticas de la física teórica nos dejen entrever en cierta medida las auténticas estructuras generales de la realidad. Según este punto de vista, podría haber algo de cierto en la tesis pitagórica de «las matemáticas, esencia de lo real». Contrariamente a los cubos de las ruedas (por volver a la comparación esbozada antes), los «números» de la física podrían ser en parte no creaciones nuestras sino, más incluso que el «lugar» y la «cosa», elementos de la propia realidad. No obstante, esta visión, tan incierta como grandiosa, no debería inducirnos a menospreciar el papel, esencial e innegable, de las «matemáticas no ontológicas»: los «números de la física», forjados por nosotros como el carretero forja los cubos de las ruedas, que ordenan las apariencias.

En 1996, la brillante denuncia por parte de Alan Sokal de los daños causados por el sociologismo ha reavivado indirectamente el debate sobre los fundamentos de la física. Esta denuncia, urgente, fue llevada a cabo con mano maestra. Y es bueno que, una vez levantado, este gran viento barredor de nubes haga vacilar también tales o cuales atrevidas extrapolaciones filosóficas de ciertos esquemas teóricos tampoco muy firmes. La palabra «deriva» está en este caso justificada. Lamentablemente, este exigente deseo de racionalidad incita a veces a quienes anima a atrincherarse en posiciones demasiado estrechas.

BERNARD D'ESPAGNAT antiguo director del laboratorio de física teórica de partículas elementales de la Universidad de Orsay.

*GRUPO DE TRANSFORMACIONES
En matemáticas, una transformación es una función del espacio en sí mismo. Un grupo de transformaciones es un conjunto de transformaciones provisto de una estructura de grupo por una ley de composición interna.

(1) W. Heiserberg, *La Partie et le Tout*, Albin Michel, 1972.
(2) J.S. Bell, *Lo decible y lo indecible en mecánica cuántica*, Alianza Editorial, 1990.
(3) M. Gell-Mann, *El quark y el jaguar*, Tusquets, 1995; véase *Mundo científico*, diciembre de 1995.

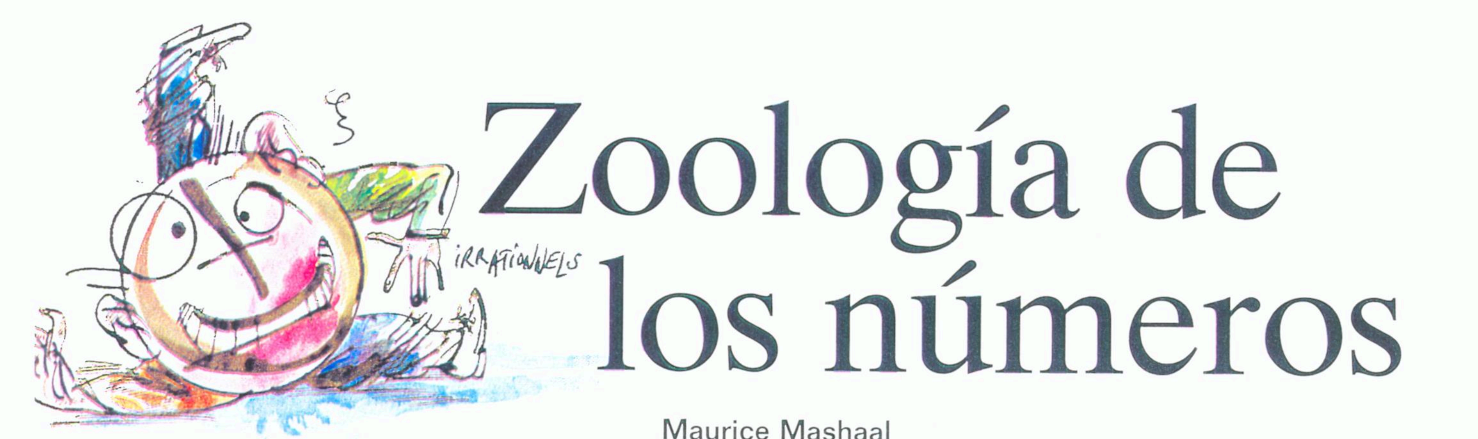
Mundo Científico ha publicado:
(I) «Las soprendentes predicciones de la mecánica cuántica», enero de 1987.
(II) «Una nueva interpretación de la mecánica cuántica», diciembre de 1995.

PARA MÁS INFORMACIÓN:

- Bernard d'Espagnat y Etienne Klein, *Regards sur la matière, des quanta et des choses*, Fayard.
- Bernard d'Espagnat, *Ondine et les feux du savoir*, Stock, 1998.

Más detallado (y que por lo tanto exige una lectura más atenta):

- John Allen Paulus, Érase una vez un número, Tusquets Editores, Barcelona, 1999.



Zoología de los números

Maurice Mashaal

Los números enteros, naturales y negativos, los números racionales, los números irracionales, los números reales, los números complejos, etcétera. Sería imposible realizar un inventario completo de los diferentes tipos de números que han construido los matemáticos. He aquí los más utilizados y algunos ejemplos.

ENTEROS NATURALES

Se llaman así los números enteros positivos 1, 2, 3, 4,... El conjunto de estos números se denota **N***. Cuando incluye el cero se denota **N**.

NÚMEROS PRIMOS

Un número entero positivo se dice primo si sólo es divisible por sí mismo y por 1 (véase el artículo de Henri Cohen en este número). Por ejemplo 3, 5, 67, 103 son números primos. Todo entero positivo se puede escribir de forma única como un producto de números primos.
Por ejemplo: $504 = 2^3 \times 3^2 \times 7$.

NÚMEROS DE FERMAT

Son los números de la forma $2^{a(n)} + 1$ con $a(n) = 2^n$. Para $n = 0, 1, 2, 3, 4$ estos números son primos. Pierre de Fermat los creía equivocadamente primos para todo n; de hecho no se conoce ningún otro número de este tipo que sea primo. A la edad de 19 años, Carl Friedrich Gauss demostró que un polígono regular que tenga un número primo p de lados se puede construir con regla y compás si y sólo si p es un número de Fermat primo.

NÚMEROS DE MERSENNE

Son números enteros de la forma $M_p = 2^p - 1$. Si M_p es un número primo, el entero p es necesariamente primo. Hasta ahora sólo se conocen 33 números de Mersenne primos, el mayor de los cuales (de momento...) fue descubierto en 1994 y corresponde a $p = 859.433$. No se sabe si existe o no una infinidad de números de Mersenne primos.

NÚMEROS PRIMOS GEMELOS

Se trata de pares de números primos cuya diferencia es igual a 2, como (5, 7) y (17, 19).

NÚMEROS PERFECTOS

Un número entero positivo se dice perfecto si es igual a la suma de sus divisores (excepto él mismo, pero incluyendo 1).
Los tres primeros números perfectos son 6, 28 y 496:
$6 = 1 + 2 + 3$,
$28 = 1 + 2 + 4 + 7 + 14$
$496 = 1 + 2 + 4 + 8 + 16 + 31 + 62 + 124 + 248$.
Euclides, en el siglo III antes de nuestra era, demostró que un número de la forma $2^{p-1}(2^p-1)$ es perfecto si 2^p-1 es primo.
En el siglo XVIII, Leonhard Euler demostró que estos números son los únicos números perfectos pares. En cuanto a los números perfectos impares, no se sabe si existen (se cree que no).

NÚMEROS AMIGOS

Dos enteros positivos m y n se dicen amigos si la suma de los divisores de m es igual a n y simultáneamente la suma de los divisores de n es igual a m.
Los números amigos más pequeños son:
220 (divisores: 1, 2, 4, 5, 10, 11, 20, 22, 44, 55, 110) y 284 (divisores: 1, 2, 4, 71, 142).

NÚMEROS PITAGÓRICOS

Tres enteros positivos a, b, c son pitagóricos si verifican la ecuación de Pitágoras $a^2 + b^2 = c^2$.

ENTEROS

Cuando se añade un signo «menos» a los enteros naturales, se obtienen los enteros negativos (-1, -2, -3, etc.). Junto con los enteros naturales, constituyen el conjunto de los enteros (..., -3, -2, -1, 0, 1, 2, 3,...). Este conjunto, también llamado de los «enteros relativos» se denota **Z**.

NÚMEROS RACIONALES

Se trata de los números que se pueden escribir como un cociente de dos enteros, es decir, como una fracción.
En otras palabras, un número racional se puede escribir en la forma p/q, donde p y q son enteros (la escritura no es única). Por ejemplo 1/2, -2/3, 8/9 = 16/18, 6/2 = 3/1 = 3 son números racionales.
El conjunto de los números racionales se denota **Q**. El conjunto **Z** de los enteros es un subconjunto del mismo.
El desarrollo decimal de un número racional contiene siempre un grupo de cifras (eventualmente una sola cifra) que se repite indefinidamente a partir de cierta posición. Por ejemplo 23,0215454545454... es el desarrollo decimal de 253.237/11.000.

NÚMEROS IRRACIONALES

Son los números cuyo desarrollo decimal es infinito y no periódico. Por lo tanto no se pueden escribir como un cociente entre enteros. Por ejemplo, $\sqrt{2} = 1,4142...$ es irracional, como demostraron los pitagóricos, Otros ejemplos de irracionales: (π, $\sqrt{5}$, $1 + \sqrt{3}$, etc. De hecho, la mayoría de los números son irracionales. En efecto,

su infinidad es de un orden superior a la infinidad de los números racionales.[1] Además la mayoría de los racionales también son trascendentes (véase más adelante).

NÚMEROS REALES

El conjunto de los números reales agrupa al conjunto de los racionales y al conjunto de los irracionales; se denota **R**.

El conjunto **R** fue rigurosamente «construido» durante la segunda mitad del siglo XIX por diversos métodos. Contribuyeron varios matemáticos (el francés Charles Méray, los alemanes Karl Weierstrass, Richard Dedekind y Georg Cantor, etc.). En el método de Cantor, se consideran sucesiones de elementos de **Q** (es decir de números racionales) que poseen una determinada propiedad técnica (se trata de las «sucesiones de Cauchy»). Un ejemplo lo constituye la sucesión de números racionales 1, 1 – 1/3, 1 – 1/3 + 1/5, 1 – 1/3 + 1/5 – 1/7, etc. Esta sucesión converge hacia el número 0,78539816... (que es igual a ($\pi/4$. El límite de este tipo de sucesiones de racionales no es siempre un número racional, como lo demuestra este ejemplo. Pero con estos límites se pueden definir las mismas operaciones (suma, multiplicación, etc.) que en los números racionales. En este caso, el conjunto de los números reales se obtiene «completando» el conjunto **Q** con los límites de sucesiones convergentes de números racionales. Así construido, el conjunto **R** comprende los números racionales y los números irracionales (que por medio de esta construcción, están bien definidos), y además permite describir todos los puntos de una recta: a todo punto de una recta se le puede asociar un número real y sólo uno, y recíprocamente.

NÚMEROS COMPLEJOS

Estos números fueron inventados en el siglo XVI por los italianos Jerónimo Cardan y Rafaello Bombelli especialmente, con objeto de resolver ecuaciones que no tienen solución en números reales, como $x^2 + 1 = 0$ y $x^4 + 2 = 0$.

La idea consiste en introducir un símbolo *i* (de *imaginario*) que verifica $i^2 = -1$ (es decir, que formalmente se puede escribir $i = \sqrt{-1}$; la notación *i* fue introducida en 1777 por Euler). Este símbolo «resuelve» por lo tanto la ecuación $x^2 + 1 = 0$, imposible de resolver anteriormente. Se extiende entonces la noción de número considerando todas las combinaciones posibles de la forma $a + bi$, donde a y b son números reales. El conjunto de estos números, llamados complejos, se denota **C**. Incluye los números reales (caso $b = 0$). Los números complejos cuya parte real a es nula (como el número $2i$) se llaman *imaginarios*.

El interés de los

infinitamente más «numerosos» que los enteros y los racionales (para más detalles y un desarrollo de los ordinales transfinitos, véase).

ALGUNOS NÚMEROS NOTABLES

π

El número π (pi) es sin duda la constante más célebre de las matemáticas. Es el cociente entre el perímetro de la circunferencia y su diámetro, y vale 3,1415926535... Desde la antigüedad existen aproximaciones a este cociente.

El francés Jean Henri Lambert demostró en 1768 que π es irracional. En 1882, el alemán Ferdinand von Lindemann demostró que es trascendente, es decir no algebraico. Con ello demostró, al mismo tiempo, la imposibilidad de la cuadratura del círculo: no se puede, con la ayuda tan sólo de regla y compás, construir un cuadrado cuya área sea exactamente igual a la de un círculo (ya que esta construcción se puede traducir en forma de una ecuación algebraica de la que π tendría que ser solución).

Existe un gran número de expresiones de π en forma de series infinitas. Por ejemplo, $\pi/4 = 1 - 1/3 + 1/5 - 1/7 + 1/9 + ...$, y también $\pi^2/6 = 1 + 1/22 + 1/32 + 1/42 + ...$

e

Este número igual a 2,7182818284... desempeña un papel por lo menos tan importante como el número π. Es el límite de $(1 + 1/n)^n$ cuando n tiende a infinito. El número e, así designado por Euler en 1728, es la base de los logaritmos neperianos (o logaritmos naturales): la notación y = ln x significa que $e^y = x$.

Se puede demostrar que $e^x = 1 + x + x^2/2! + x^3/3! + ...$ (donde, por definición, $n! = n(n-1)(n-2)...4.3.2.1$).

Se tiene así $e = 1 + 1 + 1/2! + 1/3! + ...$ Fue partiendo de este desarrollo que Euler demostró que e es irracional. El francés Charles Hermite demostró en 1973 que e es trascendente.



La función exponencial e^x desempeña un papel fundamental en análisis. Por ejemplo, está ligada a funciones trigonométricas por la relación $e^{ia} = \cos a + i \operatorname{sen} a$ (donde $i = \sqrt{-1}$), que se utiliza muy frecuentemente para facilitar los cálculos trigonométricos. Esta fórmula permite además obtener $e^{i\pi} = -1$.

Otro ejemplo que ilustra su importancia: la derivada de la función e^x es igual a sí misma. Ésta es la razón por la que la función exponencial interviene de forma esencial en la resolución de ecuaciones diferenciales lineales.

CONSTANTE DE EULER

Denotada (por Euler, es el límite de $1 + 1/2 + 1/3 + ... + 1/n - \ln n$ cuando n tiende a infinito. Esta constante vale 0,5772156649.... Euler calculó los 16 primeros decimales de γ. Todavía ni se sabe si este número es irracional, y todavía menos si es trascendente.

NÚMERO ÁUREO

Representado con frecuencia (τ o ϕ), mide la antigua «sección áurea» de los griegos. En un segmento AB, la sección áurea viene determinada por un punto M tal que AB/AM = AM/MB. Cada uno de estos cocientes es entonces igual al número áureo, y vale $(1+\sqrt{5})/2$, es decir 1,61803398.... (para algunos autores, el número áureo designa al inverso de este número, es decir: $(\sqrt{5}-1)/2$ = 0,61803398...).

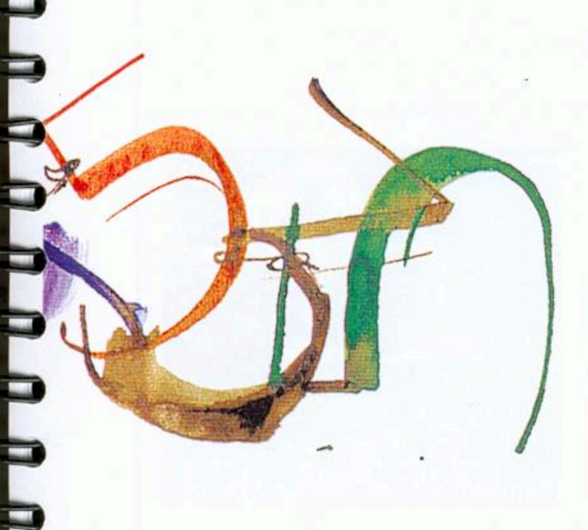
Se trata de un número irracional, pero algebraico, solución positiva de la ecuación $x^2 - x - 1 = 0$. El número áureo también está relacionado con la sucesión de Fibonacci 1, 1, 2, 3, 5, 8, 13, ..., en la que cada término es la suma de los dos precedentes. Se demuestra en efecto que el cociente entre el $(n+1)$ —ésimo término y el enésimo término —es decir el cociente entre dos elementos consecutivos de esta sucesión— tiende hacia el número áureo cuando n tiende a infinito.

El número áureo, también llamado divina proporción, ha desempeñado un papel en la estética clásica (véase el artículo de Marguerite Neveux en este número). Se le atribuyen interpretaciones místicas y simbólicas.

Más pertinente desde el punto de vista científico es el hecho de que el número áureo aparece en filotaxia, es decir en la disposición de las hojas alrededor del tallo en las plantas.(II)

MAURICE MASHAAL es periodista de *La Recherche*.

Mundo Científico ha publicado:
(I) H. Sinaceur, «El infinito», nº151, noviembre 1994.
(II) S. Douady, Y. Couder, «La física de las espirales vegetales», nº 133, marzo, 1993.



PARA MÁS INFORMACIÓN:

F. Le Lionnais (con la colaboración de J. Brette), *Les nombres remarquables*, Hermann, 1983 y 1994
D. Wells, *The Penguin dictionary of curious and interesting numbers*, Peguin Books, 1987
E. Maor, *e: the story of a number*, Priceton University Press, 1994
A. Bouvier, M. George y F. Le Lionnais, *Dictionnaire des mathématiques*, Press Universitaires de France, 1993.
A.Warusfel, *Les nombres et leurs mystères*, Seuil, 1961.
A. Dahan -Dalmedico y J. Peiffer, *Une histoire des mathématiques*, Seuil, 1986.

Buscar, jugar, encontrar

Elisabeth Busser, Francis Casiro, Gilles Cohen y Benoît Rittaud



COSAS DE NÚMEROS

«Dios hizo los enteros, el resto es obra de los hombres», dijo el matemático Kronecker. El resto es enorme: los reales, los racionales, los irracionales los «trascendentes» hasta los que no son reales llamados números complejos. Toda una fauna se ha reunido aquí para jugar con vosotros y relataros sus aventuras.

Números enteros

1 = 2

Los matemáticos han rivalizado en ingenio para encontrar (falsas) pruebas de esta sorprendente igualdad. He aquí algunas de estas presuntas demostraciones. ¿Sabríais encontrar sus fallos? El de la primera pseudojustificación es bastante evidente, el de la última en realidad es para especialistas.

a. Si x = 1 e y = 2, está claro que $x^2 + y = x + xy$ (=3). De esta última igualdad se deduce $x^2 - x = xy - y$, que se escribe $x(x - 1) = y\,(x - 1)$, es decir, x = y o también 1 = 2.

b. Consideremos la serie:
$2 - 1 + 1 - 1 + 1 - 1 + \ldots$
Utilizando el sistema de paréntesis siguiente: $(2 - 1) + (1 - 1) + (1 - 1) + \ldots$, su suma vale 1. Si, en cambio, los paréntesis están dispuestos así: $2 + (-1 + 1) + (-1 + 1) + \ldots$,
la suma vale 2. Leibniz aseguraba incluso que se podía atribuir el valor 3/2 a la suma de la serie basándose en el desarrollo: $1/(1 + x) = 1 - x + x^2 - x^3 + x^4 - x^5 + \ldots$ y tomando x = 1.

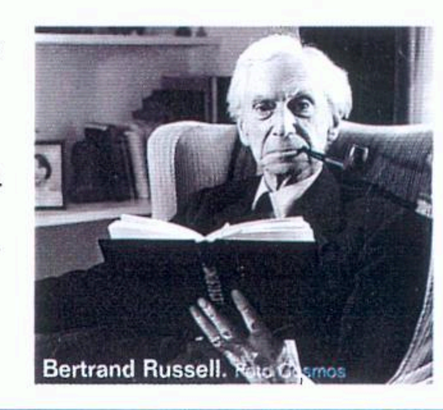
c. Por medio de matemáticas todavía más sofisticadas:
Sabemos (si ln designa el logaritmo neperiano) que $\ln(1 + x) = x - (1/2)x^2 + (1/3)x^3 - (1/4)x^4 + \ldots$ para $-1 < x \leq 1$.
Sea x = 1. $\ln 2 = 1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + 1/7 - 1/8 + 1/9 - \ldots$
Multipliquemos los dos miembros por 2: $2 \ln 2 = 2 - 2/2 + 2/3 - 2/4 + 2/5 - 2/6 + 2/7 - 2/8 + 2/9 - \ldots = 2 - 1 + 2/3 - 1/2 + 2/5 - 1/3 + 2/7 - 1/4 + 2/9 - \ldots$
Agrupemos los términos que tienen el mismo denominador, $2 \ln 2 = 1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + 1/7 - 1/6 + \ldots = \ln 2$.
Finalmente, 2 = 1.

2

La numeración usual es la numeración decimal, en la que las diez cifras 0, 1, 2, ...9 permiten representar todos los números. Pero la numeración binaria, que ha experimentado un renacimiento desde la aparición de los ordenadores, consigue representar todos los números con sólo 0 y 1. La última cifra representa las unida-

2 + 2 = 5

¿Quién no se habrá sentido sorprendido, en un primer momento, por el principio de la lógica según el cual de una proposición falsa puede deducirse cualquier otra? Así le ocurrió a un estudiante de filosofía, quien buscó en Russell alguna aclaración:
«¿Pretende que de 2 + 2 = 5 se deduce que usted es el papa?»
«Sí», dijo Russell.
«¿Y podría demostrarlo?», preguntó el estudiante escéptico.
«Ciertamente», replicó Russell, quien propuso de inmediato la demostración siguiente.
(1) Supongamos que 2 + 2 = 5
(2) Restemos 2 de cada miembro de la identidad, con lo que obtenemos 2 = 3.
(3) Por simetría, 3 = 2.
(4) Restando 1 de cada miembro, 2 = 1.
Ahora bien, el papa y yo somos dos. Como 2 = 1, el papa y yo somos uno. Por tanto, soy el papa.



Bertrand Russell. Foto Cosmos

des (1 para un número impar, 0 para un número par), el anterior las «dosenas», el anterior las «cuatrenas», el anterior las «ochenas», etc. La numeración binaria está en la base de la curiosa manera de multiplicar siguiente. Estudiémosla en un ejemplo, el producto de 457 por 321. La idea es muy simple. Se encabezan dos columnas con los números a multiplicar y luego, en cada una de las líneas siguientes, se divide por dos el número de arriba en la primera columna (sin preocuparse del resto) y se multiplica por dos el número de arriba en la segunda. Se termina cuando el cociente (primera columna) es 1. Esto es:

457	321
228	642
114	1 284
57	2 568
28	5 136
14	10 272
7	20 544
3	41 088
1	82 176

Basta ahora sumar en la segunda columna los números (en rojo) que corresponden a un número impar de la primera columna: 321 + 2568 + 20544 + 41088 + 82176 =146697. Comprobad que tal es efectivamente el número buscado.

Explicad el mecanismo de esta multiplicación

3

Reflexión de un matemático: *«Hay tres tipos de matemáticos, los que saben contar y los que no».*

5 & 17

El matemático inglés J.J. Sylvester (1814-1897) gustaba de proponer pequeños problemas en el periódico recreativo *Educational Times*. He aquí uno de ellos. Tengo un gran número de sellos cuyos valores son únicamente 5 p o 17 p.

¿Cuál es el mayor franqueo entero que se puede obtener por medio de combinaciones de estos dos valores de sellos?

10

He aquí diez proposiciones numeradas del 1 al 10. ¿Cuántas de ellas son verdaderas?

1.
Exactamente una de estas proposiciones es falsa.
2.
Exactamente dos de estas proposiciones son falsas.
3.
Exactamente tres de estas proposiciones son falsas
5.
Exactamente cinco de estas proposiciones son falsas.
6.
Exactamente seis de estas proposiciones son falsas.
7.
Exactamente siete de estas proposiciones son falsas.
8.
Exactamente ocho de estas proposiciones son falsas.
9.
Exactamente nueve de estas proposiciones son falsas.
10.
Exactamente diez de estas proposiciones son falsas.

24

El número 24 es tal que la suma de los cuadrados de los 24 primeros enteros es un cuadrado:

$1^2 + 2^2 + 3^2 + ... + 24^2 = 70^2$

Sólo otro entero verifica la misma propiedad. ¿Cuál?



J.J. Sylvester. (Foto Mary Evans/explorer)

1999

La propiedad siguiente no es en absoluto característica de 1999: todo número entero natural puede escribirse como suma de cuatro cuadrados. Se sospechaba desde hacía tiempo, pero fue Lagrange quien lo demostró en 1770.

Hallar una descomposición de 1999 en suma de cuatro cuadrados.

¿Es verdad que todo número entero se escribe como suma de cuatro cubos?

— Es verdad para ciertos enteros de la forma 9n + 4 o 9n –4.

— Se ha demostrado que es verdad para todo entero.

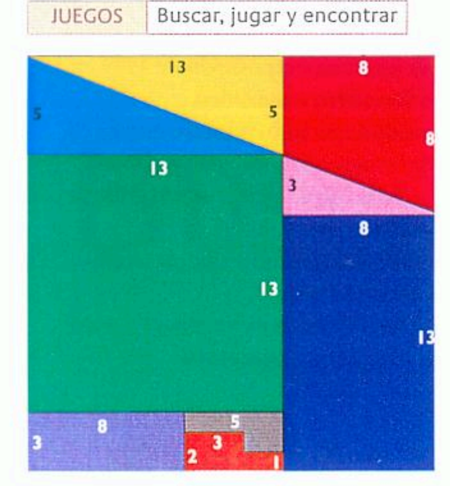
— Probablemente es verdadero pero sigue siendo una conjetura



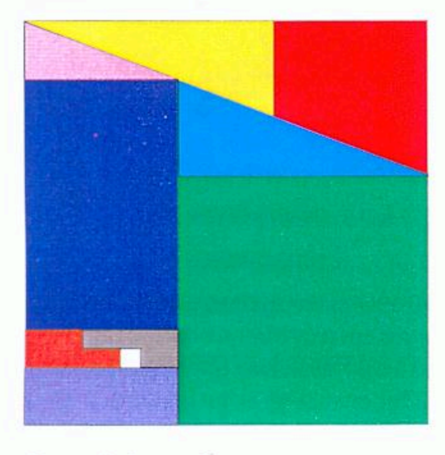
Los números de Fibonacci

Para hacerse un nombre en el universo de los enteros hay que comportarse como los payasos. Hay que sorprender. A tal fin, cuando se es un número, hay que poseer unas propiedades extraordinarias, unas propiedades que llaman la atención de generaciones de matemáticos o que son tan espectaculares que incluso llaman la atención del gran público. Entonces, es la gloria, se aparece en todos los libros, en todos los diccionarios, ¡se habla por fin de uno! Sólo que para llegar hasta aquí hace falta un mánager, a poder ser matemático. Los números de Fibonacci forman parte de este grupo.

Si os dicen 1, 1, 2, 3, 5, 8, 13... es en ellos en quienes pensáis inmediatamente. En su origen hay un asunto de conejos, estas parejas de encantadores roedores que dan origen a una nueva pareja en la generación siguiente, y a otra en la siguiente, cada una de las cuales da a su vez origen a..., etc. Es así como se construye, generación tras generación, una sucesión de números enteros que ha inspirado a los matemáticos como pocas lo han hecho. De esta sucesión se han descubierto cientos de propiedades, algunas de lo más insólito. Tomad por ejemplo 1, 2, 3, 5, 8 y 13 y utilizadlos en las dimensiones de las piezas de un rompecabezas cuadrado de 21 de lado, de la siguiente forma:



Reorganizáis los pedazos de la manera siguiente y queda un agujero. ¿Extraño, no?



¡Buscad el error!

El juego de Oslo

En el juego de Oslo se trata de obtener cualquier entero natural no nulo, partiendo del 4, por medio de aplicaciones sucesivas de una de las reglas siguientes: 1. Poner un 0 al final del número (esto es, multiplicar por 10); 2. Poner un 4 al final del número; esto es, multiplicar por 10 y añadir 4; 3 Dividir por 2 si el número es par. Es bastante fácil obtener los 10 primeros enteros positivos (en lo que sigue el símbolo Ø indica que se ha utilizado una de las tres reglas anteriores).

4 Ø 2 Ø 1 ; 4 Ø 2; 4 Ø 2 Ø 24 Ø 12 Ø 6 Ø 3; 4; 4 Ø 2 Ø 1 Ø 10 Ø 5.

Completad la sucesión hasta obtener los 10 primeros números enteros. En vuestra opinión, ¿es posible engendrar así todos los enteros naturales no nulos?

Sucesiones lógicas

Los tests de inteligencia incluyen a menudo preguntas donde, dada una sucesión de números, se pide adivinar el término siguiente. Por ejemplo, ¿cuál es el decimotercer término de la sucesión que empieza por: 1, 2, 4, 8, 16, 31, 57, 99, 163, 256, 386, 562? El cálculo de la tabla de diferencias permite brillar por poco precio (pero ¡cuidado! ¡no siempre es aplicable!) Se trata, dada una línea de números, de buscar las diferencias entre un término y el precedente a fin de obtener una nueva línea y de volver a empezar hasta que sólo figuren números idénticos.

Así, con la sucesión propuesta, se obtiene:

Valores iniciales:

1 2 4 8 16 31 57 99 163 256 386 562

Primeras diferencias

1 2 4 8 15 26 42 64 93 130 176

Segundas diferencias

1 2 4 7 11 16 22 29 37 46

Terceras diferencias

1 2 3 4 5 6 7 8 9

Cuartas diferencias

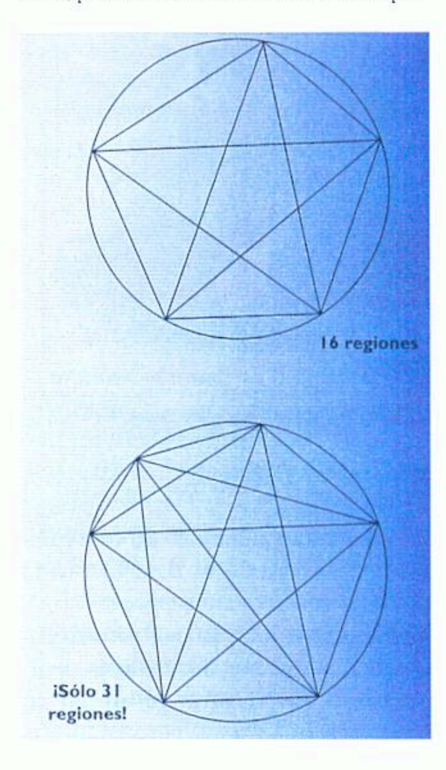
1 1 1 1 1 1 1 1

Conjeturando que la regularidad constatada se perpetúa, basta remontarse en la tabla desde la última línea después de haber añadido un «1»...

Se obtiene así, sucesivamente: 0, 56, 232 y 794. El resultado buscado es 794.

La sucesión propuesta responde a la pregunta siguiente: ¿cuál es el número máximo de regiones que se obtienen en un disco por medio de cuerdas que unen n puntos de todas las maneras posibles?

Se observa que nos encontramos ante un ejemplo típico de falsa inducción, pues si nos limitamos a los 5 primeros términos tenemos la impresión de que se trata de la serie de potencias de 2, lo que resulta inexacto.



Otro problema:

Si los valores iniciales representan las potencias *n*-ésimas de los enteros positivos sucesivos, ¿qué se lee en la línea de las n-ésimas diferencias?

En cambio, las diferencias sucesivas no os serán de ninguna ayuda para resolver esta célebre «sucesión automática».

He aquí los 5 primeros términos de la sucesión. ¿Cuál es el término siguiente?

0

10

1110

3110

132110

El algoritmo de Göbel

He aquí el algoritmo ideado por un matemático llamado Göbel. Dado un número de rango 1, los números de los rangos siguientes se calculan con la ayuda del siguiente algoritmo recurrente:

- Entrad el número N de rango R. Haced el producto de N por (N + R)
- Dividid este producto por (N + 1)
- El resultado es el número de rango (R + 1).

Si el número de rango 1 vale 2, **calculad los números de rangos 2, 3, 4, y 5.** Se trata de números enteros.

¿Creéis que a todo rango le corresponde un número entero?

Elegid una de las respuestas siguientes:

— A todo rango le corresponde un número entero; está demostrado.

— Se conjetura que a todo rango le corresponde un número entero, pero nadie lo ha demostrado.

— El número de rango 43 no es entero.

Racionales y algebraicos

Fracciones egipcias

Los antiguos egipcios conocían los números fraccionarios, pero, curiosamente, se limitaban a los inversos de los enteros naturales no nulos. Una fracción egipcia es un número de la forma $1/n$, donde n es un entero. En una tablilla de arcilla, se encontró el problema siguiente: «¿Cómo repartir equitativamente siete panes entre diez hombres?». La respuesta propuesta por el escriba no es «siete dé-

cimos de pan», sino « la mitad de un pan más un quinto de un pan». Toda fracción p/q puede escribirse de modo evidente como suma de fracciones egipcias:

1/q + 1/q + ... (1/q repetida p veces). Pero los escribas, además, se prohibían repetir la fracción. ¿Puede escribirse sin embargo toda fracción como suma de fracciones egipcias si se prohíbe la repetición de la misma fracción? Leonardo de Pisa (apodado Fibonacci, 1175-1250), propuso en su *Liber abaci* un algoritmo que respondía al problema: «*Restar a la fracción dada la mayor fracción egipcia posible, repetir la operación con la nueva fracción y así sucesivamente hasta que la operación dé una fracción egipcia.*»

Apliquemos el algoritmo de Fibonacci a 3/25.

Se encuentra $\frac{3}{25} = \frac{1}{9} + \frac{1}{113} + \frac{1}{25425}$.

Se vé así que el algoritmo de Fibonacci presenta el defecto de no dar siempre la descomposición más simple; así, $\frac{3}{25} = \frac{1}{10} + \frac{1}{50}$ es una solución más satisfactoria.

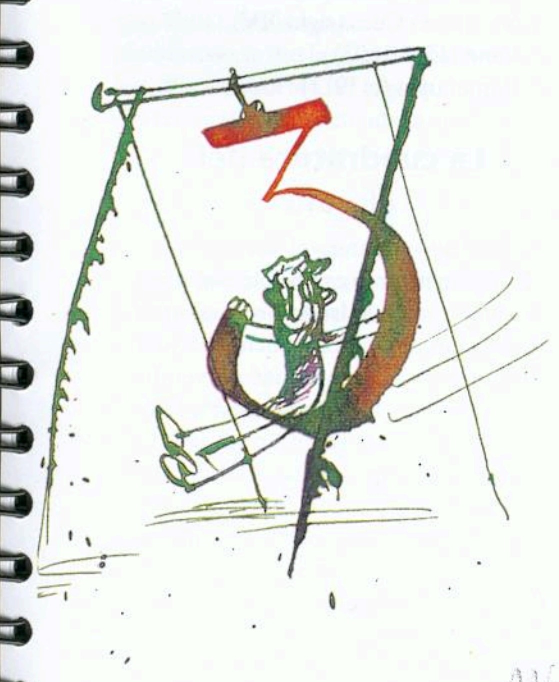
Sylvester demostró en 1880 que este algoritmo daba siempre una solución y que para p<q, p/q se escribía como suma de un máximo de p fracciones egipcias distintas.

Cuadrados de fracciones egipcias

El resultado más notable ligado a las fracciones egipcias es la fórmula de Euler:

$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} ...$

(al no ser π una fracción, la suma tiene que ser infinita)



¿Se puede escribir 1/2 como suma de cuadrados de fracciones egipcias?

Una extraña calculadora

Mi calculadora está en un lamentable estado. Las únicas teclas de operación que funcionan son +, – y 1/x, que permiten sumar, restar e invertir.

¿Puedo obtener, de todos modos, el producto de dos reales?

De ser así, ¿en cuántas operaciones como mínimo?



√2

El descubrimiento de los números irracionales se remonta a la escuela pitagórica. Aunque no sabemos cómo ocurrieron las cosas, parece ser que la diagonal fue la piedra del «escándalo lógico» (por utilizar la expresión de Paul Tannery). La diagonal de un cuadrado de lado 1 mide √2: los griegos decían que la diagonal realizaba la duplicación del cuadrado (el área del cuadrado construido sobre la diagonal es el doble de la del cuadrado inicial). ¿Dónde está, pues, el escándalo? En que √2 no se expresa como cociente entre dos números enteros, en que no es una fracción.

Por ello, para los pitagóricos la medida de la diagonal carecía de sentido. Las distintas denominaciones dadas a este tipo de números («*absurdos*», «*irregulares*») indican bastante bien su estado de ánimo. Actualmente, cuando ya han dejado de plantear problemas, se les llama «*irracionales*». Aristóteles cita a menudo en sus escritos el caso de la diagonal, «*inconmensurable*» al lado del cuadrado (aunque en su época la teoría de los irracionales estaba ya bien elaborada).

√3

Es posible que los griegos no se hubieran dado tanta prisa en zanjar la crisis de los irracionales de no haber aparecido estos de un modo tan natural en geometría. Así ocurre con la diagonal de un rectángulo de lados 1 y 2, de longitud √5 según el teorema de Pitágoras.

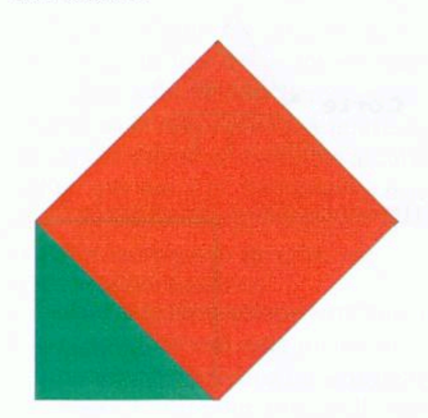
Dad dos métodos geométricos simples para hacer aparecer √3.

√2 + √3

La secta de los pitagóricos no sobrevivió al descubrimiento de los números irracionales. Pero de entre estos irracionales, todavía hoy sorprendentes, podemos sentirnos familiarizados con los números «algebraicos», no necesariamente racionales pero raíces (soluciones) de una ecuación polinómica de coeficientes racionales. Son algebraicos √2 y √3, por supuesto (son raíces de los polinomios $x^2 - 2$ y $x^2 - 3$), pero ¿sabéis que toda suma de números algebraicos es algebraica? La demostración general es difícil, pero **¿sabríais encontrar un polinomio del que √2 + √3 sea raíz?**

$\sqrt{2}^{\sqrt{2}}$

Por medio del número A = $\sqrt{2}^{\sqrt{2}}$ ¿podríais demostrar simplemente que existe un irracional que elevado a una potencia irracional da un número racional?

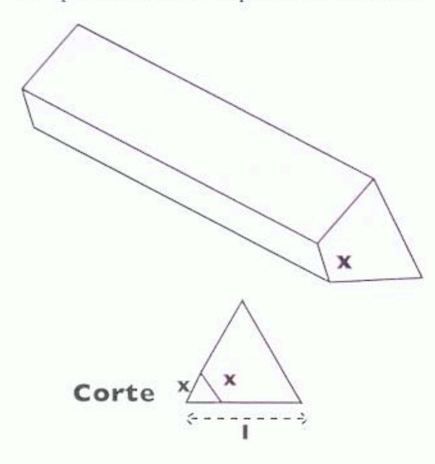


Φ

Conocido con el nombre de divina proporción y luego de número áureo desde Luca Pacioli, que le dedicó la obra de su vida, *Divina Proportione*, Φ, que vale $\frac{1+\sqrt{5}}{2}$, se ha visto atribuir todas las virtudes, por lo que al hablar de este número se tiene la molesta impresión de que todo ha sido dicho

acerca de él. Como se sabe, sirve para dividir un segmento en extrema y media razón, determina la razón entre la diagonal y el lado de un pentágono regular y se le asocia generalmente el rectángulo y el triángulo áureos, cuyos lados están en esta proporción. También es, dentro de otras muchas propiedades, el límite del cociente de dos números consecutivos de la sucesión de Fibonacci.

Por dos razones, el número áureo es sinónimo de equilibrio. En primer lugar porque un rectángulo áureo siempre ha significado para los artistas una forma bien proporcionada, y luego porque este número tiene efectivamente algo que ver con el equilibrio de los cuerpos. Colocad sobre una de sus caras rectangulares un prisma recto cuya base sea un triángulo equilátero. Cortadle un «rincón», como en la figura. Si la parte quitada es lo bastante pequeña, todo va bien, pero cortad un poco más, y más aún ... y el prisma acaba desplomándose. Suponiendo que el lado del triángulo sea igual a 1 y la longitud del corte en la base igual a x, ¿a partir de qué valor de x se produce la caída?

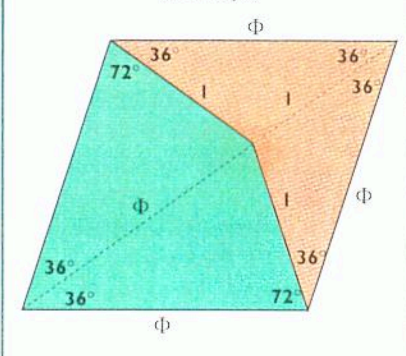


Números trascendentes

π

Los números algebraicos, que incluyen los racionales, son extremadamente raros en comparación con los demás, llamados números trascendentes. De estos números conocéis al menos uno, con el que os torturan desde la escuela primaria, el famoso número π. Esta razón *«mágica»* no siempre ha tenido nombre. Ni siquiera fue reconocido como número en la remota Antigüedad, sino como una aproximación que permitía obtener el área de un círculo conociendo su diámetro. Se le habría podido llamar número de Arquímedes, pero no fue así.

También podéis encontrar el número áureo en el fondo de vuestra sartén. Los nuevos revestimientos antiadhesivos de las cacerolas utilizan unos recubrimientos no periódicos descubiertos por Penrose. Los motivos se construyen a partir de un rombo y Conway los ha llamado «flecha» y «cometa». Los identificaréis fácilmente en el dibujo.



¿Sabríais encontrar al menos una manera de recubrir el plano con estas dos formas de losas (dejando aparte la colocación borde a borde de los rombos)?

Los alemanes le dan a veces otro nombre, el de Ludolff von Ceulen, un matemático que ya en 1596 había calculado 20 decimales exactos y luego 34 en 1609. La notación π (como περιμετρου, perimetron) fue ideada por Jones en 1706 pero no se impuso hasta 1748 con Euler (el cual, mientras tanto, había llamado p a la constante, mientras que Bernoulli la había bautizado c).



Las aproximaciones de π

En el Antiguo Testamento (*Libro de los reyes*, VII, 23) se encuentra ya esta famosa razón constante entre el perímetro de una circunferencia y su diámetro. Como podréis constatar, vale π = 3.

«Hizo el mar de bronce fundido. Tenía diez codos de un borde al otro; era completamente redonda; su altura era de cinco codos y un cordón de treinta codos medía su circunferencia.»

El primer teorema serio sobre el valor de π fue enunciado por Arquímedes: *«El área de un círculo es al cuadrado de su diámetro aproximadamente como 11 es a 14. El perímetro de todo círculo es mayor que el triple del diámetro y excede este valor en una longitud inferior a la séptima parte del diámetro y superior a diez setentaiunavos.»*. Históricamente, se han hecho otras aproximaciones. He aquí, al hilo de los siglos, algunos de los valores atribuidos a π. **¿Sabríais situar cada uno de ellos en su contexto histórico?**

Lista de valores aproximados

a. $3\frac{1}{8}$

b. 377/120

c. $\left(102-\dfrac{2222}{22^2}\right)^{\frac{1}{4}}$

d $\dfrac{2^8}{3^4}$

e. $3/4\,(\sqrt{3}+\sqrt{6})$

f. $2\times\dfrac{2}{\sqrt{2}}\times\dfrac{2}{\sqrt{2+\sqrt{2}}}\times\dfrac{2}{\sqrt{2+\sqrt{2+\sqrt{2}}}}\times..$

(la primera fórmula «infinita» de π)

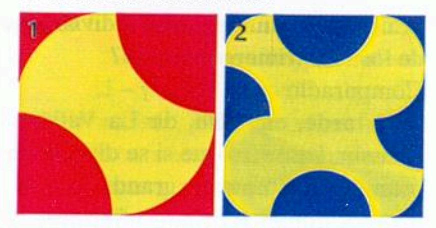
Lista de los conceptos históricos:

1. Los babilonios (hace 4000 años)
2. El Papyrus Rhindt (1800 a.C.) descubierto en 1855
3. Tolomeo (c. 100 d.C.)
4. Nicolás de Cues (siglo XV)
5. Viète (1540-1603)
6. Ramanujan (c. 1913)

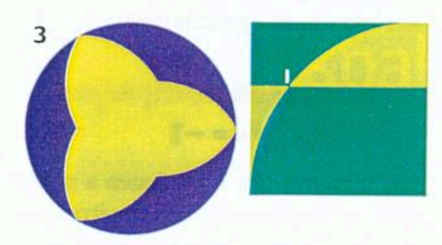
La cuadratura del círculo

La idea de «cuadrar» el círculo, es decir, de construir, si es posible con regla y compás, un cuadrado de igual área que un círculo dado, preocupaba ya a los antiguos. Durante más de veintitrés siglos, este famoso problema de la cuadratura del círculo ha impedido dormir a más de un matemático. Hipócrates de Quios, en el siglo V a.C., había logrado cuadrar algunas «lúnulas», figuras formadas por arcos de círculo. Lo intentaron después muchos geómetras, entre los cuales Leonardo da Vinci, quien según se dice dibujó un centenar. Todos ellos tuvie-

ron la impresión de progresar hacia la resolución de un problema histórico. La cuestión fue definitivamente zanjada en 1882 cuando Lindemann, al probar la trascendencia de π, demostró también la imposibilidad de la cuadratura del círculo.



¿Sabríais, sin ningún cálculo, determinar el área de las tres lúnulas de arriba?



En este cuadrado, que suponemos de lado 1, inscribimos el cuarto de círculo dibujado, que cortamos en I por medio de una paralela a los lados horizontales.

¿Para qué posición del punto I es mínima el área de amarillo?

e

Más joven que π, el otro número faro de los trascendentes, introducido casi artesanalmente en el siglo XVI por John Napier en relación con los logaritmos, pasó casi desapercibido hasta que Euler, adivinando su futuro esplendoroso, le dio la inicial de su nombre. Para nosotros, el logaritmo son dos teclas de una calculadora, ln y log, pero la invención de este concepto fue crucial para los astrónomos de los siglos XVI y XVII, ya que les permitió sustituir sus largas multiplicaciones por sumas. En una palabra, el logaritmo permite asociar a la sucesión de potencias sucesivas de un mismo número (la base) la sucesión de los enteros.



Potencias de 10	**1**	**10**	**100**	**1000**
Logaritmo decimal (de base 10)	**0**	**1**	**2**	**3**
Potencias de 2	**0**	**2**	**4**	**8**
Logaritmo de base 2	**0**	**1**	**2**	**3**

El logaritmo inventado por Napier, que llamamos neperiano, tiene por base un número extraño, llamado e por Euler, que vale alrededor de 2,718281828... El número e, como π, tiene un interés (en un sentido muy literal, como veréis) que rebasa el marco de los cálculos logarítmicos. Por ejemplo, invertís una modesta suma, digamos que de 1 unidad, de manera particularmente ventajosa, puesto que la tasa nominal es del 100%. Permite a vuestro capital doblar todos los años si los intereses se pagan a fin de año. Pero capitalizando cada seis meses los intereses (50%), el capital se multiplica al cabo de un año por 2,25. Reduzcamos más aún el período de capitalización. Con vuestra calculadora favorita, completad el cuadro siguiente:

Período de capitalización	Capital al final del año
1 año	**2**
6 meses	**$1,5^2 = 2,25$**
3 meses	
1 mes	
1 semana	
1 día	
1 hora	
1 minuto	
1 segundo	

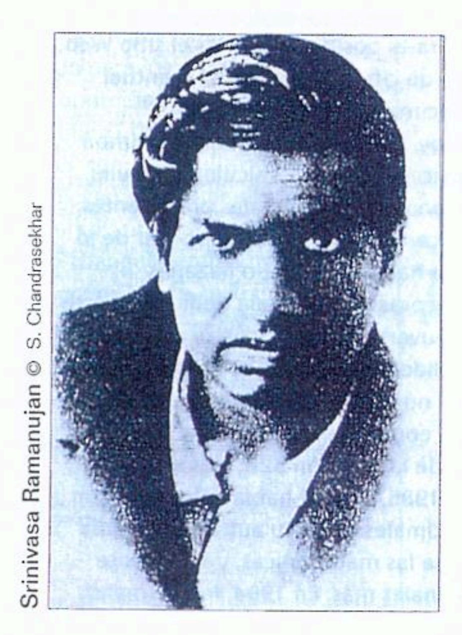
Esto es, el límite de $(1 + 1/n)^n$ cuando n tiende a infinito es precisamente e.

La Caza de los Decimales de Π

π podría figurar por más de una razón en el libro de los récords: por la larga lista de sus aproximaciones, por el número de decimales descubiertos, o por las hazañas de los prodigios capaces de memorizar interminables sucesiones decimales del número. Por ejemplo, ¿sabéis cuántos decimales de π consiguió recitar en 1979 un japonés llamado Tomoyori? ¡15 151! Y en 1995, un cierto G. Goto recitó 42 000. ¿Figura inevitablemente toda sucesión de cifras en el desarrollo decimal de π? Ha sido postulando una respuesta afirmativa a esta pregunta como ha construido Darren Aronofsky el guión de una película reciente llamada precisamente «π». Los místicos buscan en él el nombre de Dios y los financieros la clave de la Bolsa. Lo cual hace pasar malos momentos al protagonista de la película en cuestión. Ya es posible consultar el sitio Web «Am I a Pi?» para proponer una serie de cifras y conocer su eventual posición en la sucesión de decimales de π (dirección: www.facade.com/Fun/amΠnpi). Pues hoy, ¡victoria! Dabid Bailey y Simon Pouffe construyeron en 1996 un algoritmo capaz de calcular cualquier decimal de π, por lejano que sea, sin conocer los decimales precedentes. Los récords, por tanto, tienen un mero carácter anecdótico. A pesar de lo cual esta caza de decimales de π había propiciado hazañas extraordinarias. Un caso notable es la decoración de la sala dedicada a este número en 1937 en el Palais de la Découverte, en París, con ocasión de la Exposición internacional. Se había decidido decorar la sala, redonda como debe ser, con numerosos decimales del número fetiche. Pero he aquí que el calculador (Shanks en 1874) había cometido un error, que quedó inscrito en el techo de la sala a partir de la posición 528. Más tarde, los ordenadores precipitaron las cosas. En 1985, Gosper había calculado, con una fórmula genial, 17 millones de decimales de π. Su autor era el indio Ramanujan, verdadero autodidacta de las matemáticas, y su potente fórmula da, a cada nuevo paso, 8 decimales más. En 1994, los hermanos Chudnovsky, de un modo casi artesanal, con un ordenador instalado en su apartamento, lograron 4 000 millones de decimales, El último récord «homologado» tiene, que nosotros sepamos, más de 6 400 millones de decimales, aunque como hemos visto el mérito ya no es el mismo.

$e^{\pi\sqrt{163}}$

Cuál no sería la sorpresa de Srinivasa Ramanujan, prodigio indio de principios de siglo (véase *Mundo Científico* nº 181), cuando advirtió que el número $e^{\pi\sqrt{163}}$ parecía ser el entero 262 537 412 640 768 744. Haced la prueba si tenéis una calculadora o un programa capaz de manejar más de 20 cifras significativas. Ramanujan conjeturó que era un entero. La respuesta la dio en los años 1970 la revista *Scientific American*.



Srinivasa Ramanujan © S. Chandrasekhar

¡El número de Ramanujan era un entero!

Sin embargo, la información fue desmentida poco después. ¿Por qué?

— Sólo se habían calculado 2 millones de decimales, lo cual no basta.

— Había un error en la demostración.

— El número de *Scientific American* era del 1 de abril.

Trascendentes pero olvidados

Otras curiosidades pueblan el bestiario de los números y los matemáticos se complacen siempre en inventar otros nuevos.

Así, por ejemplo, los números que inventó Liouville en 1844. Se trata de sumas infinitas de C^n/k^n, donde C_n es un entero comprendido entre 0 y 9 y k un entero no nulo.

El más simple de estos números, tratado por el propio Liouville, es $\frac{1}{10^{1!}} + \frac{1}{10^{2!}} + \frac{1}{10^{3!}} \dots$, que se escribe

0,1100010000000000000000010000...

Otro número trascendente un poco olvidado, y no obstante fácil de construir, es el de Champernowne: 0,12345678910111213141516171819202122232425..., que habla por sí solo.

¿Sabríais encontrar su 2 000ª cifra decimal después de la coma?



Joseph Liouville (Foto SPL/Cosmos)

γ

La constante de Euler-Mascheroni. Calculado hasta el dieciseisavo decimal por Euler, que lo llamó γ en 1781, y atribuido más tarde a Mascheroni, este



Leonard Euler Foto Cosmos

número es uno de los más misteriosos. Todavía no sabemos si es trascendente, ni siquiera si es irracional. El logaritmo neperiano de n puede interpretarse geométricamente como el área abarcada por la hipérbola de ecuación $y = 1/x$, entre las rectas de ecuación $x = 1$ y $x = n$. Esta área es apenas menor que la de la «escalera» cuyos peldaños tienen una anchura de 1 y alturas sucesivas de 1, 1/2, 1/3, 1/4, ... es la diferencia entre estas dos áreas la que da lugar a la constante de Euler-Mascheroni:

$$\gamma = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{n} - \ln n\right)$$

No sabemos gran cosa de esta constante, cuyo valor aproximado es 0,57722. Sólo que Dirichlet demostró en 1838 que el número medio de divisores de todos los enteros de 1 a n es muy próximo a $n + 2\gamma - 1$.

¿Cuál es el número total de divisores de los 100 primeros enteros?

Comparadlo con $100 + 2\gamma - 1$.

Más tarde, en 1898, de La Vallée-Poussin demostró que si se divide un número n lo bastante grande por todos los números primos inferiores, la distancia media entre las fracciones obtenidas y el entero inmediatamente superior es... precisamente γ.

Números complejos

$e^{i\pi} = -1$

Los números e y π no tienen a priori nada en común: origen geométrico en un caso, origen algebraico en el otro, campos de aplicación diferentes. Sin embargo Euler reunió en una famosa fórmula e, la base de los logaritmos neperianos e i, símbolo de los imaginarios, el famoso número cuyo cuadrado vale –1. Se dice que desde entonces se han vuelto inseparables... ¿Podríais deducir un valor «aceptable» de i^i?

Bibliografía sucinta:

John Conway y Guy K. Richard, Le *Livre des nombres, Eyrolles*.

Guy K. Richard, *Unsolved Problems in Number Theory*, Springer.

Jean-Paul Delahaye, *Le Fascinant nombre π*, Belin.

Theoni Pappas, *Joy of Mathematics*, Tetra.

Tangente, hors-série 6, «Secrets de nombres», Archimède.

PARA MÁS INFORMACIÓN:

- Martin Gardner, *Los mágicos números del Doctor Matrix*, Editorial Gedisa, Barcelona, 2000.
- M. Gardner, *El ahorcamiento inesperado y otros entretenimientos matemáticos*, Editorial Gedisa, Barcelona, 2000.
- M. Gardner, *Festival mágico-matemático*, Alianza Editorial, Madrid, 2000.
- M. Gardner, *Carnaval matemático*, Alianza Editorial, Madrid, 2000.
- Raymond Smullyan, *Juegos de ajedrez y los misteriosos caballeros de Arabia*, Editorial Gedisa, Barcelona, 2000.
- R. Smullyan, *Juegos y problemas de ajedrez para Sherlock Holmes*, Editorial Gedisa, Barcelona, 2000.